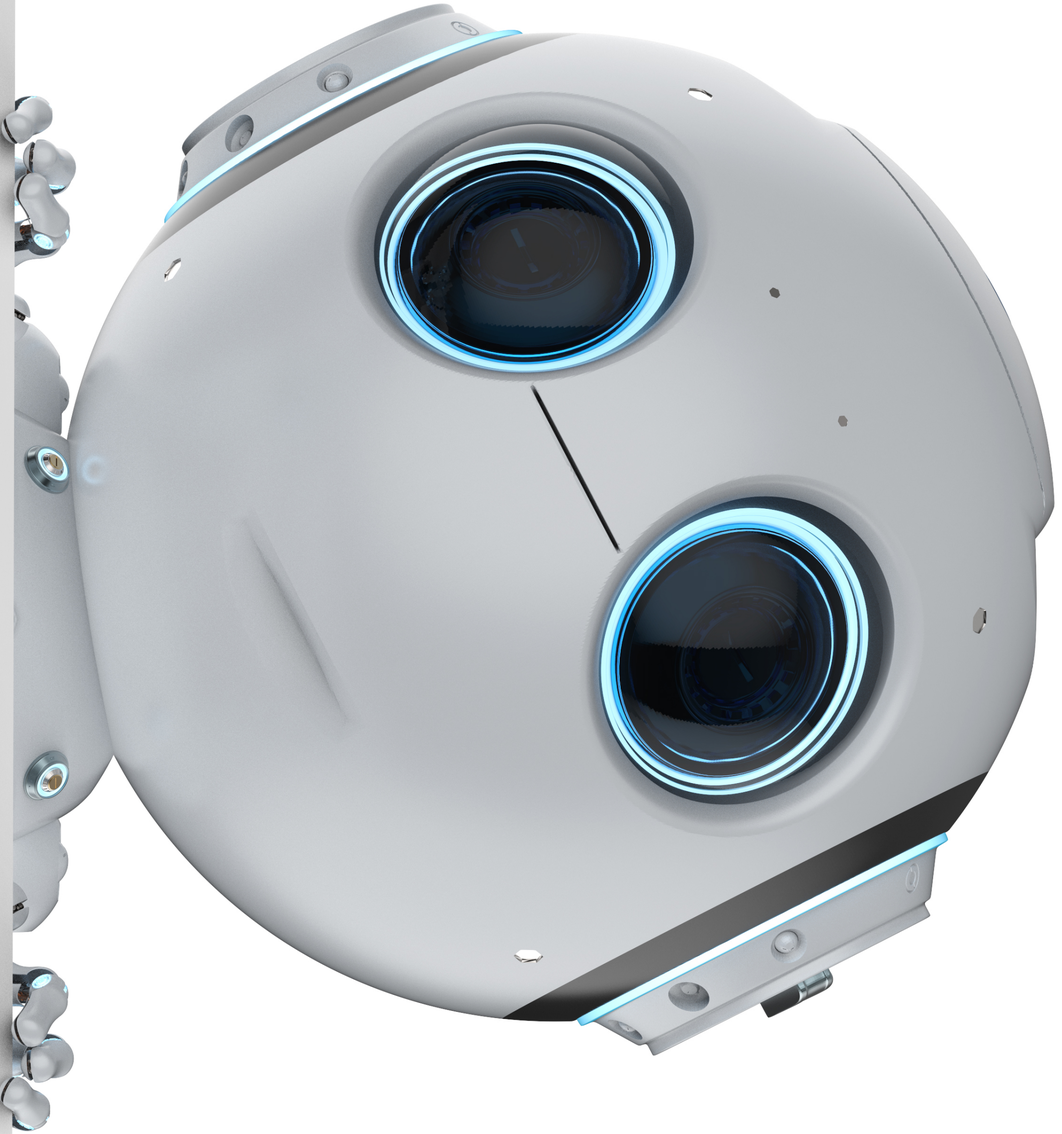
An introduction to

Responsible AI

Nina Nowak, Senior Data Scientist at Combient Mix

Nov 19, 2020



AI in Society

Opportunities: improve society

Healthcare

Accessibility (for people with disabilities)

Education

Climate and environment

Agriculture

Public services

Public safety

Home security, smart homes and cities

Automation of transport

Automation of repetitive tasks

Areas of concern: used as oppressive force

Military, police, government

Social scoring

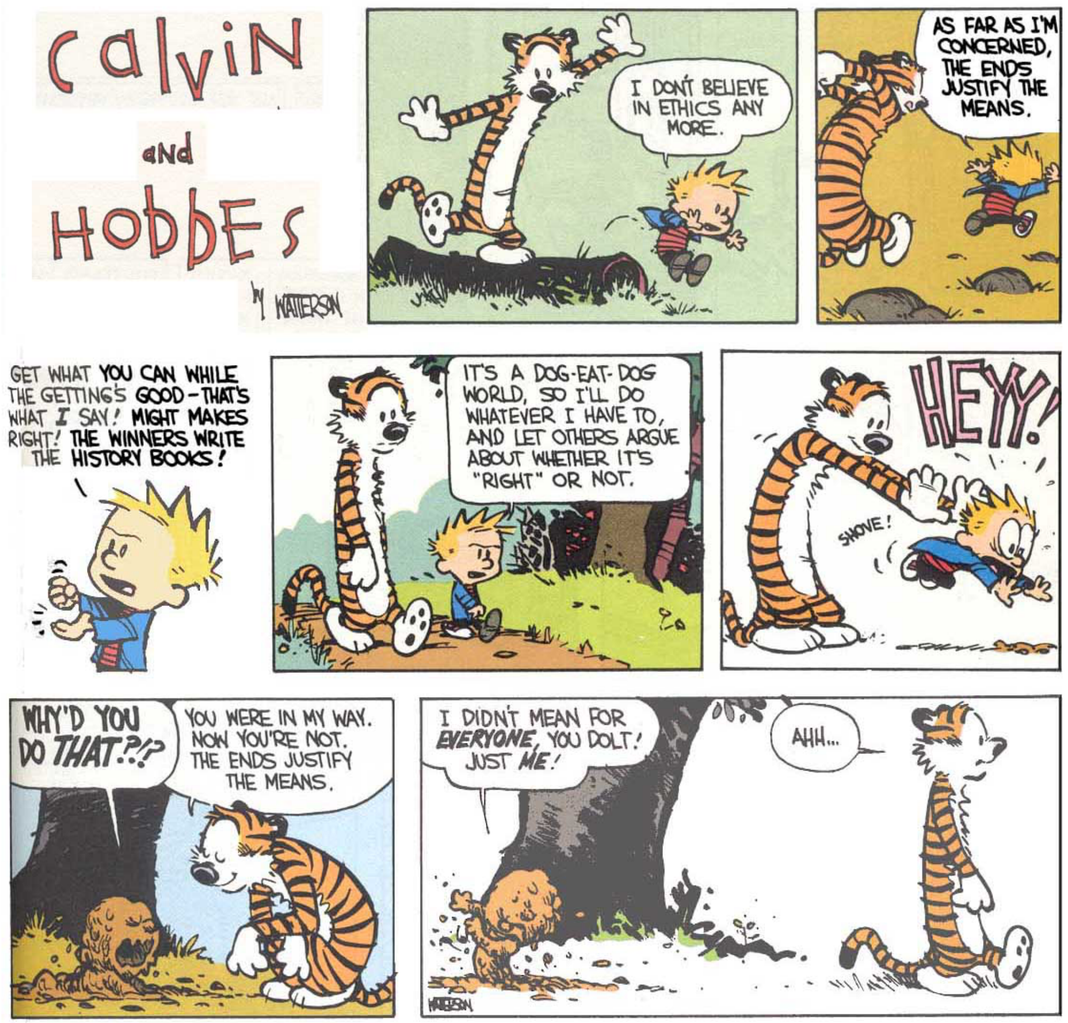
AI taking over the world

Socio-technical systems: understand the social context!

What are Ethics?

- Branch of philosophy
- Moral principles governing behaviour or actions
- Codified processes that help determine what is right or what is wrong
- Values guide ethical decision making:
  - values based on well-being of others: (non) suffering, autonomy, equality
  - values based on my own well-being: character excellence / virtue, trust
  - values are abstract guiding principles influenced by culture, socioeconomics etc.
  - ranking of values can depend upon context ➜ tradeoffs
- there are no right or simple answers, instead multiple ethical perspectives influenced by factors like experience, culture, context

Normative ethics



Moral standards that regulate right and wrong conduct
Three distinct perspectives:

Virtue Ethics
Focuses on the motivations of an action, i.e. on the inherent character of a person doing an action
In RRI: stakeholder views and values

Deontology
Morality inherent in the action. Universal ethical rules: certain things are always right or always wrong irrespective of context
E.g. it is always wrong to kill, or lie, or steal

Consequentialist and utilitarian ethics
Assess moral goodness of an action only on the basis of the outcome
You should do the thing that results in the most good or happiness for the greatest number of people
The end justifies the means

What is responsible AI?

Lawful

AI systems must acknowledge and obey fundamental rights, laws at country and multinational level

Ethical

AI systems should adhere to ethical principles and values. They should improve/ benefit individuals and society. AI systems should neither cause nor exacerbate harm. They should be designed with particular attention to vulnerable persons.

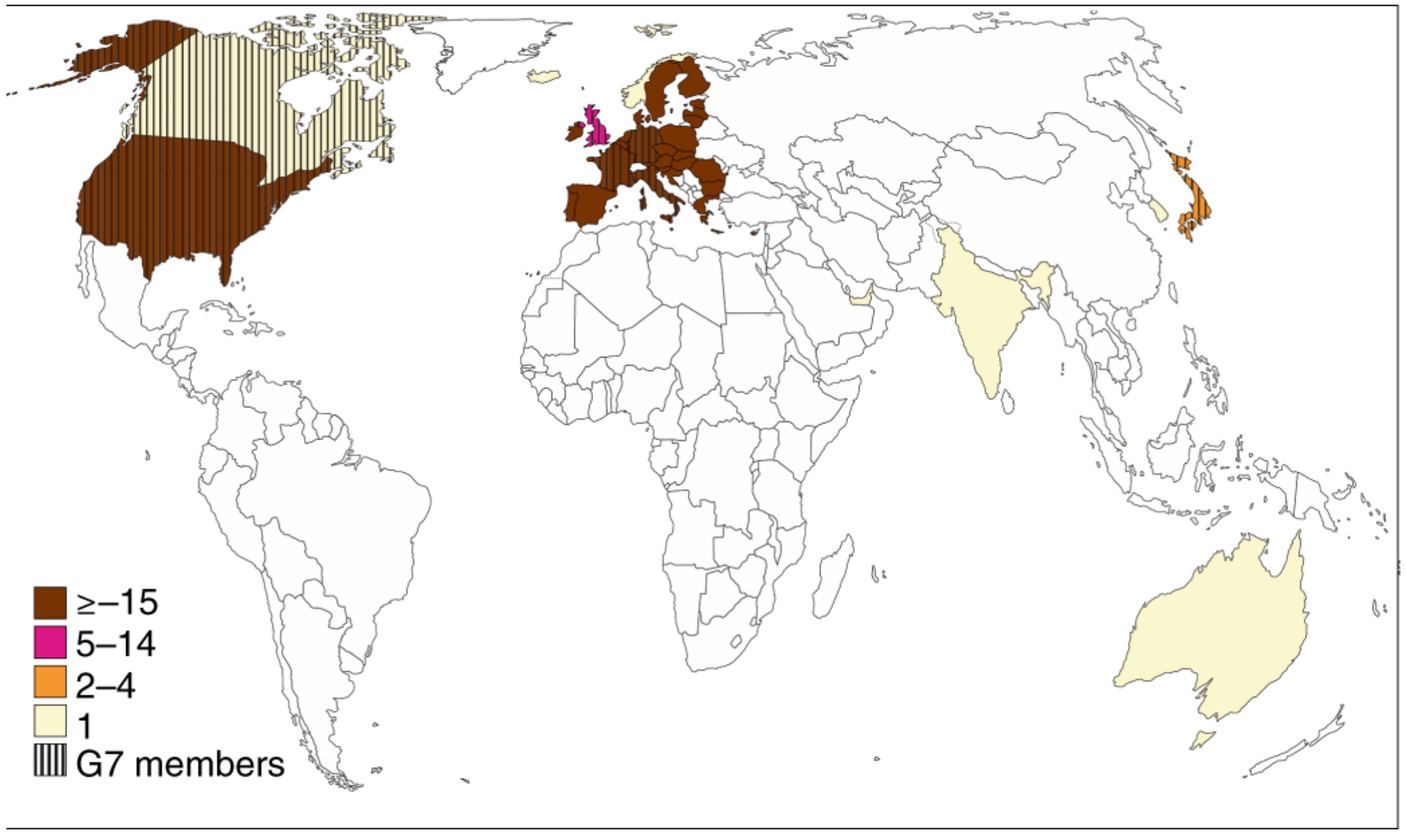
Robust

AI systems must be technically robust and it should be ensured that they are not open to malicious use. They should also be robust from a social perspective (behave as anticipated, designed so users feel safe).

.

Collection of Ethical AI guidelines

- Transparency
- Justice and fairness
- Non-maleficence
- Responsibility
- Privacy
- Beneficence
- Freedom and autonomy
- Trust
- Sustainability
- Dignity
- Solitarity



84 documents containing ethical principles or guidelines for AI (Jobin et al. 2019, Nature Machine Intelligence)

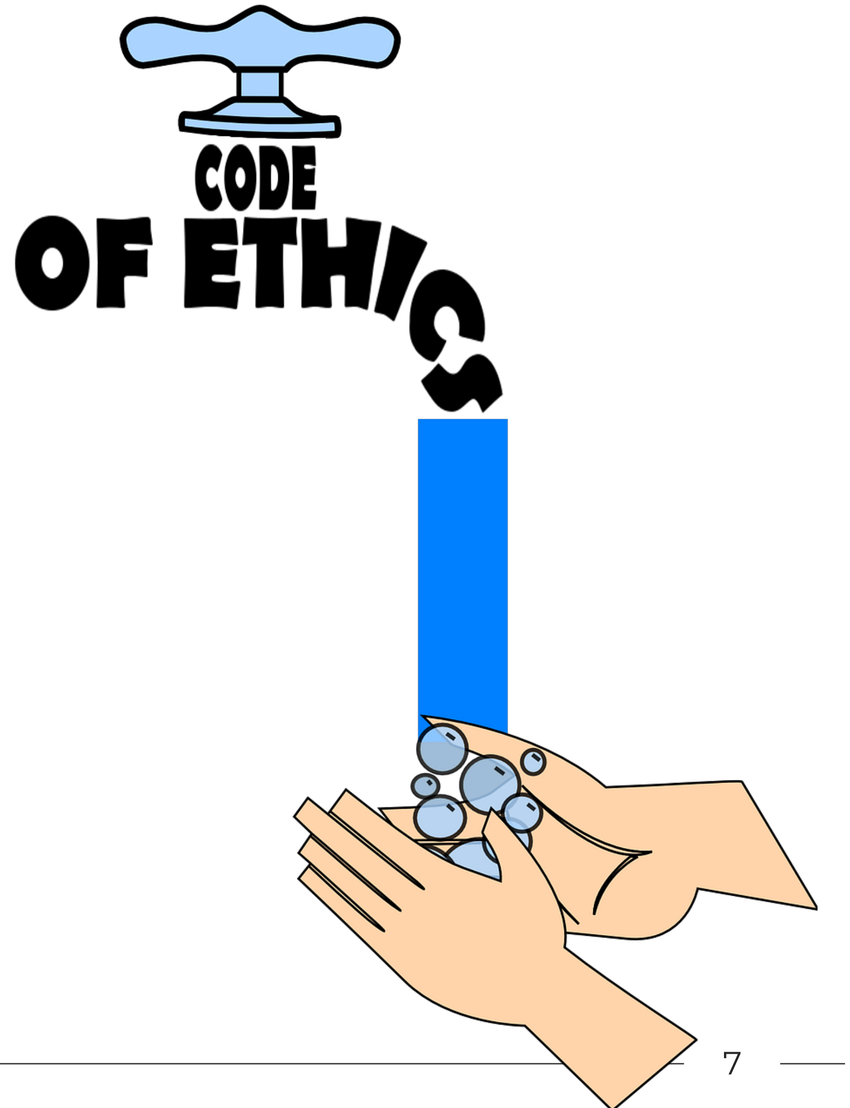
Ethics washing

Many companies create their own Code of Ethics. Valid fear of public this might
- be just a marketing tool
- create false sense of security

Impossible to ascertain if commitment is met in daily practice.

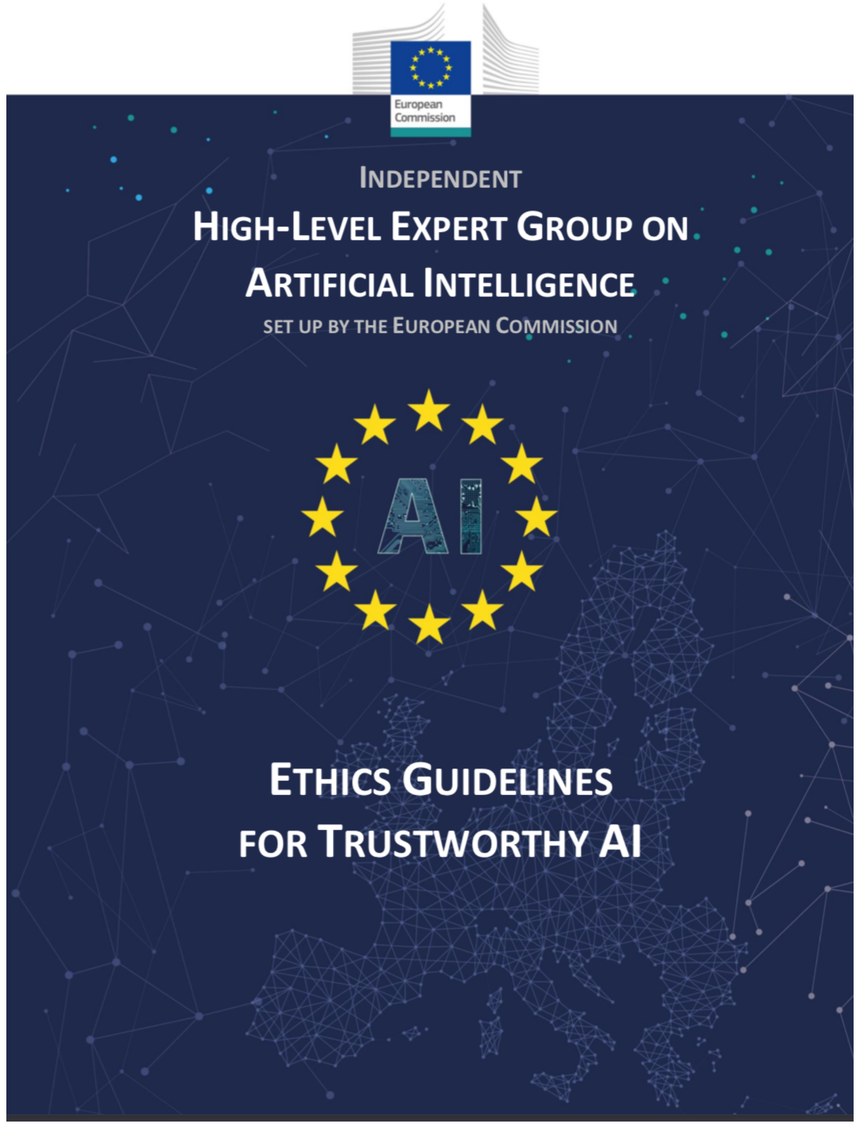
Problematic if these are just words without action, enforcement and monitoring.

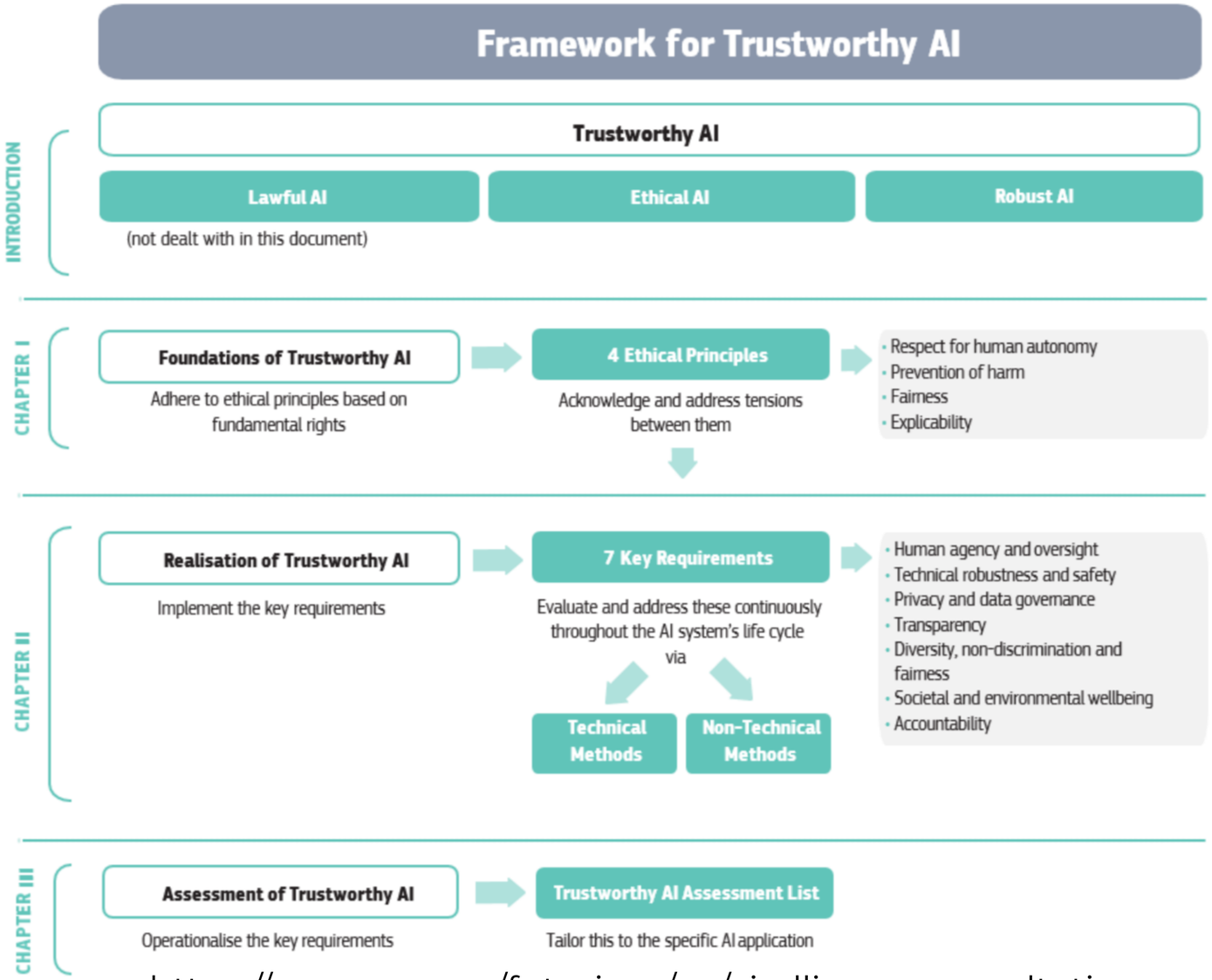
Legally binding regulations would create trust, but is slow compared to technological innovation.



Ethics guidelines for trustworthy AI

AI HLEG, European Commission





https://ec.europa.eu/futurium/en/ai-alliance-consultation


MORAL
DILEMMA
AHEAD

Tradeoffs

Not all ethical principles can be satisfied simultaneously, so tradeoffs between principles/values are common.

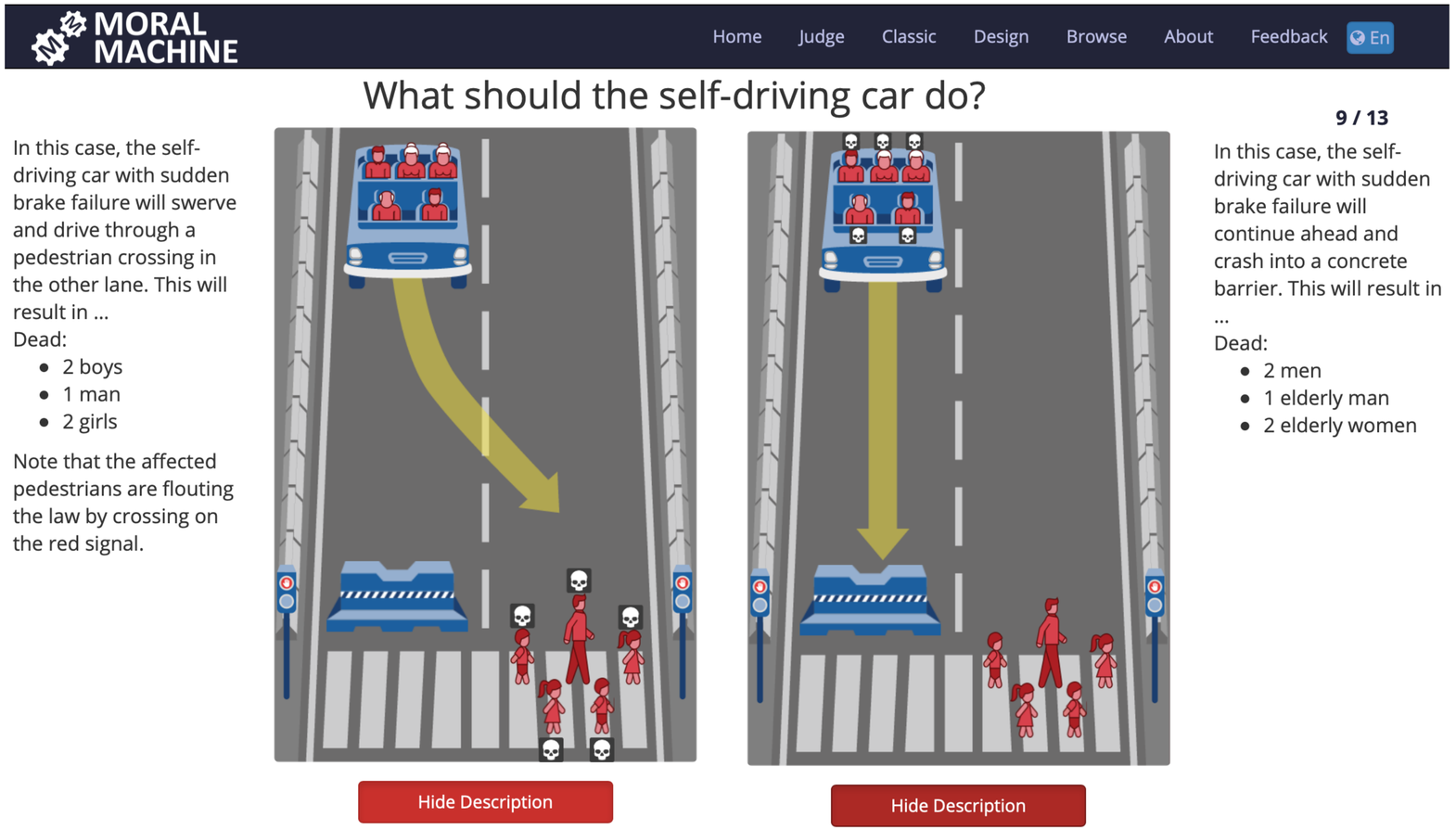
Examples:

- Surveillance cameras: Privacy vs. Safety
- Gene banks: societal benefits vs. personal cost (privacy, security)
- Hate speech on social media: freedom of speech vs. protection from harm

Tradeoffs depend on context, culture, values and prior experience. There are always positive benefits and potential negative outcomes for any solution.

Moral Machine

Moral dilemmas, where a driverless car must choose the lesser of two evils



Can’t be solved by any simple normative ethical principles

https://www.moralmachine.net

Moral Machine - Results

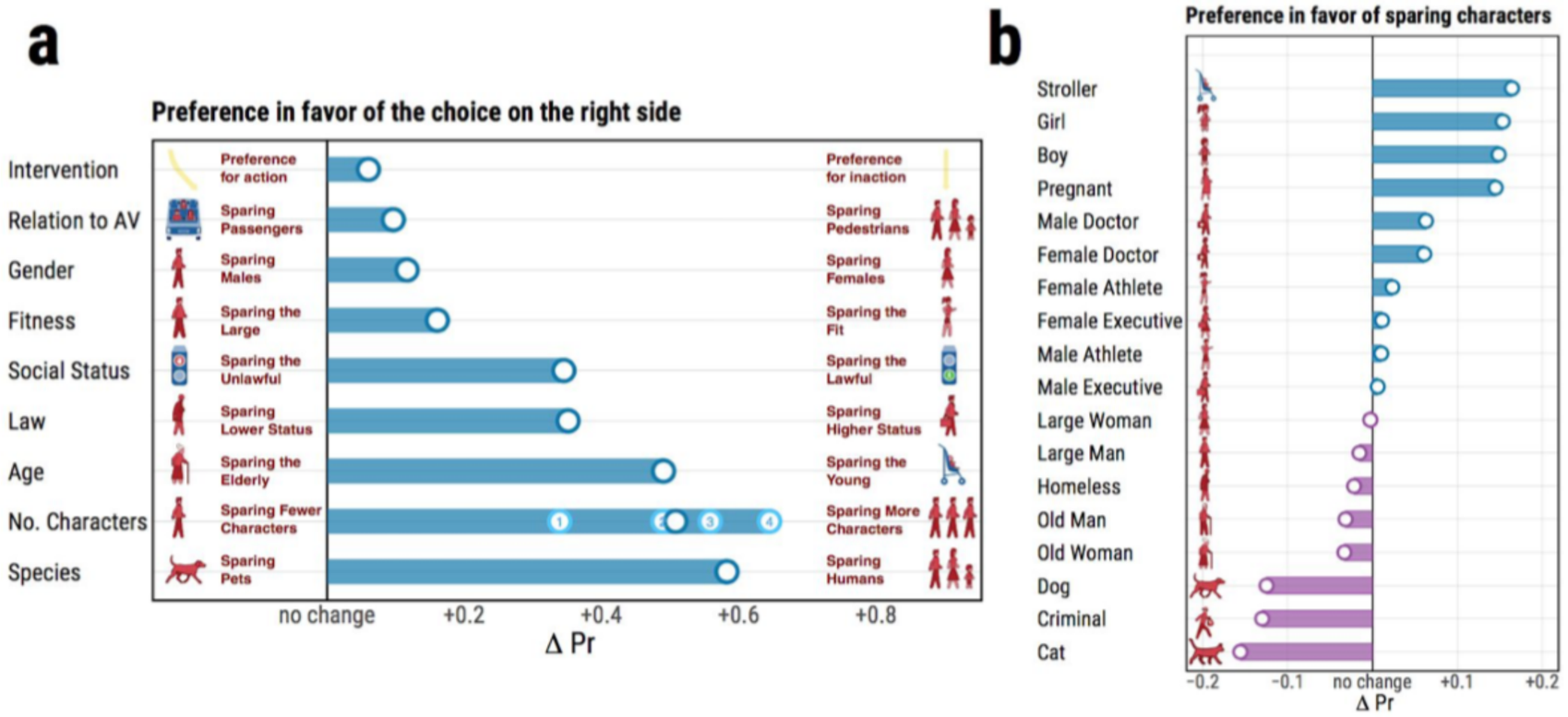
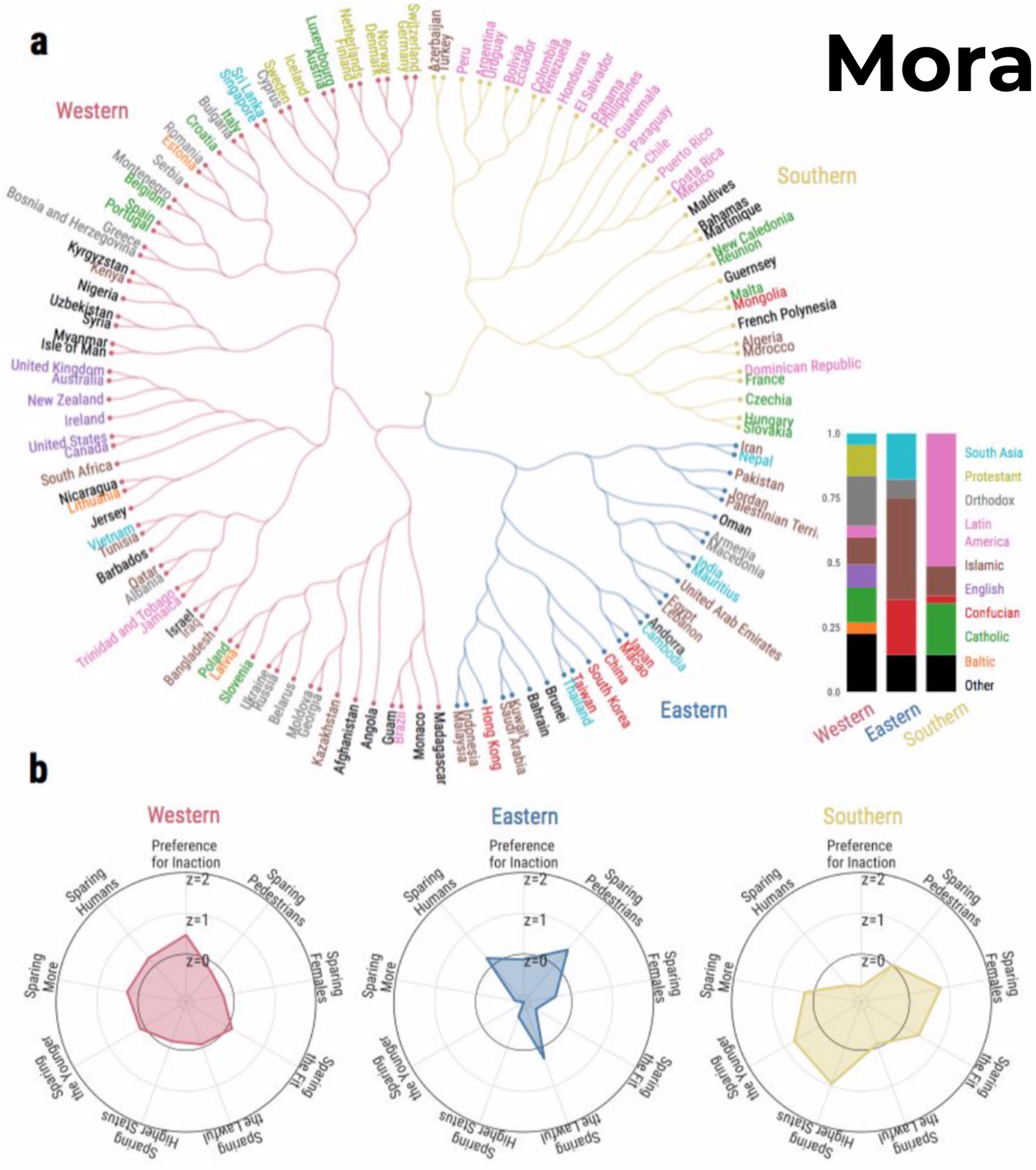


Figure 2. (a) Average marginal causal effect (AMCE) for each preference. *In each row, ΔPr is the difference between the probability of sparing characters possessing the attribute on the right, and the probability of sparing characters possessing the attribute on the left, aggregated over all other attributes. For example (age) the probability of sparing young characters is 0.49 (SE = 0.0008) greater than the probability of sparing older characters. The 95% CIs of the means are omitted due to their insignificant width, given the sample size. For the number of characters (No. characters), effect sizes are shown for each number of additional characters (1 to 4); the effect size for 2 additional characters overlaps with the mean effect of the attribute.* ***(b) Relative advantage or penalty for each character, compared to an adult man or woman.*** *For each character, ΔPr is the difference the between probability of sparing this character (when presented alone) and the probability of sparing one adult man or woman. For example, the probability of sparing a girl is 0.15 (SE = 0.003) higher than the probability of sparing an adult man/woman.*

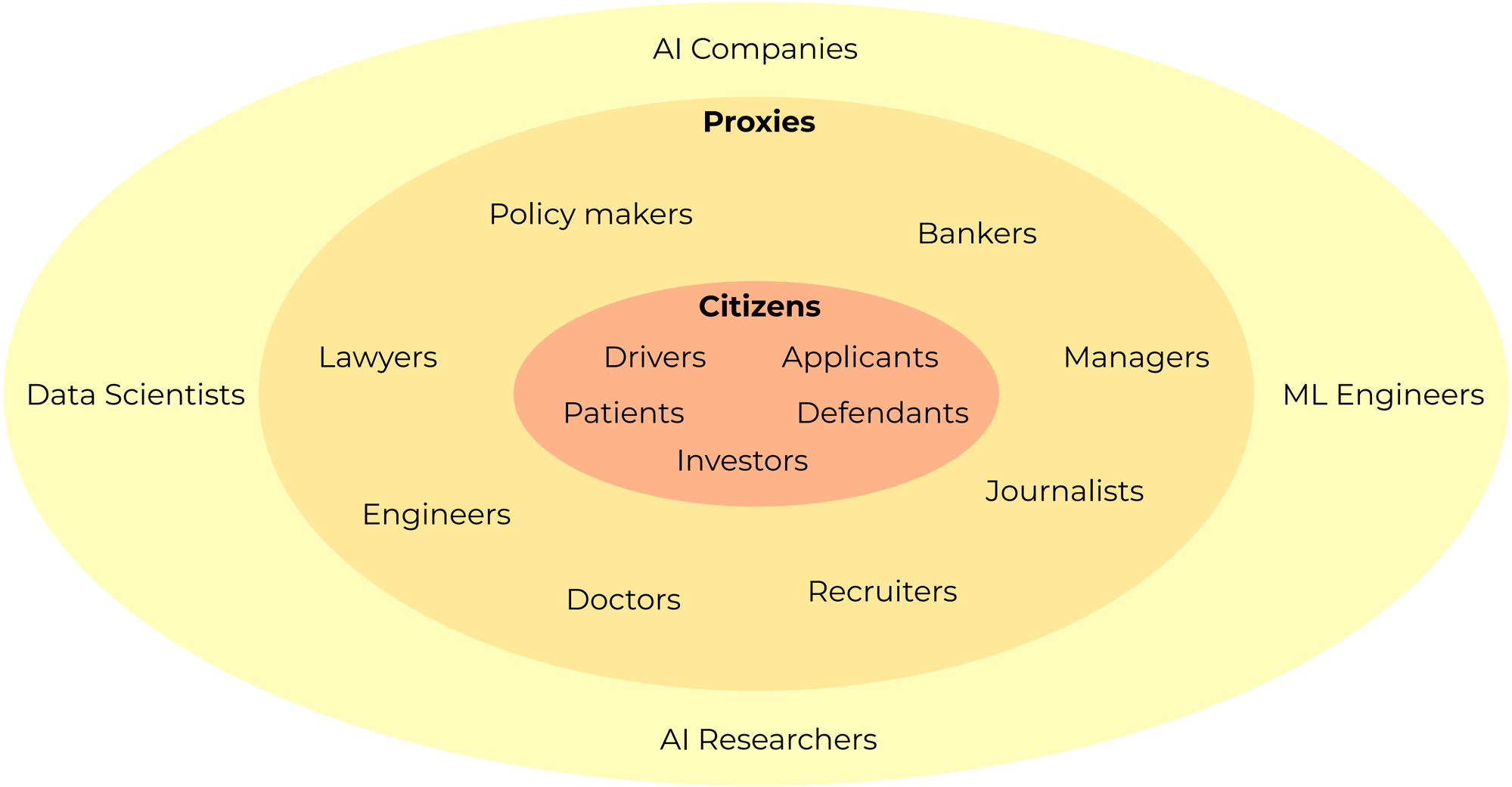
Moral Machine - Results



Hierarchical clustering:

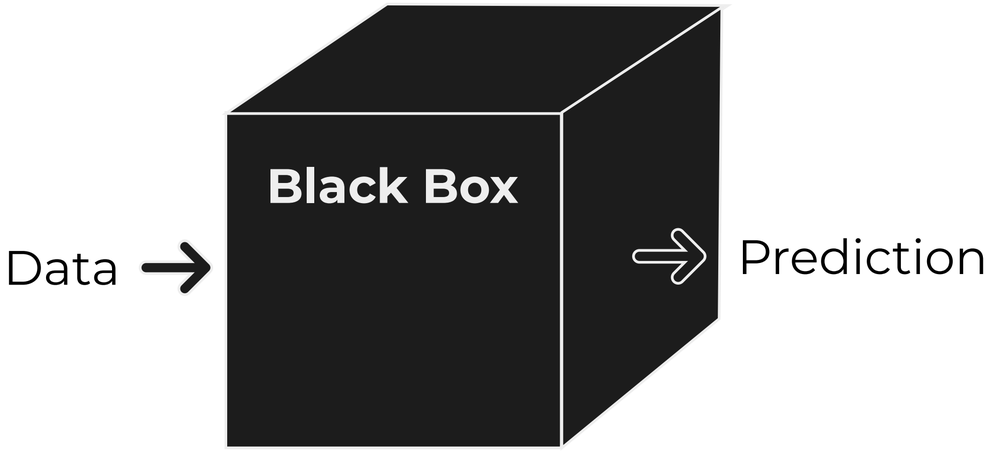
- 3 cultural clusters
- difference in weight for some preferences, e.g. spare younger characters much less pronounced in Eastern cluster
- some peculiarities like sparing women and fit in the Southern cluster
- Manufacturers of AV's should be aware of these cultural differences
- Regulators must be aware of citizen's expectations

Transparency



Transparency builds trust ➜ higher adoption rate of AI
AI systems are inherently complicated ➜ simple and efficient explanations

Transparency and Trust



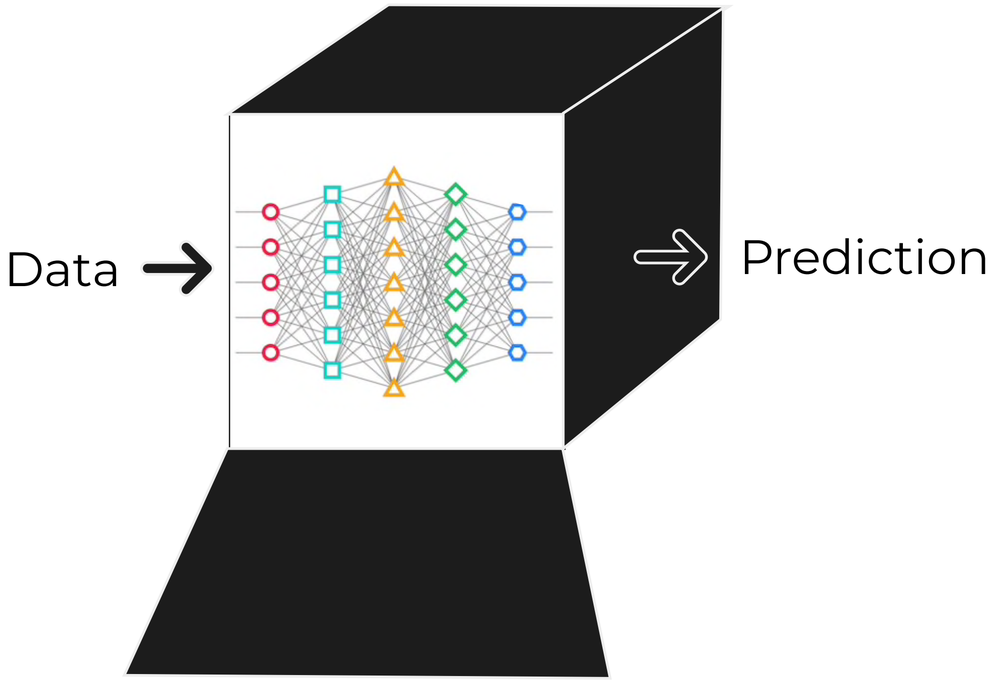
Data
What data is used, how was it collected, how is it governed, how does it fit the context of use?

Design processes
What are the assumptions, how is the data processed, what features are chosen or not chosen and why?

Algorithms
What algorithms are used, how do they come to a particular decision, for what decision criteria is the model optimised and is that justified given the context?

Stakeholders
Who is involved in the process, what are their interests, who will be affected, who are the users, is participation voluntary, who is controlling?

Visualisation as a tool to build trust



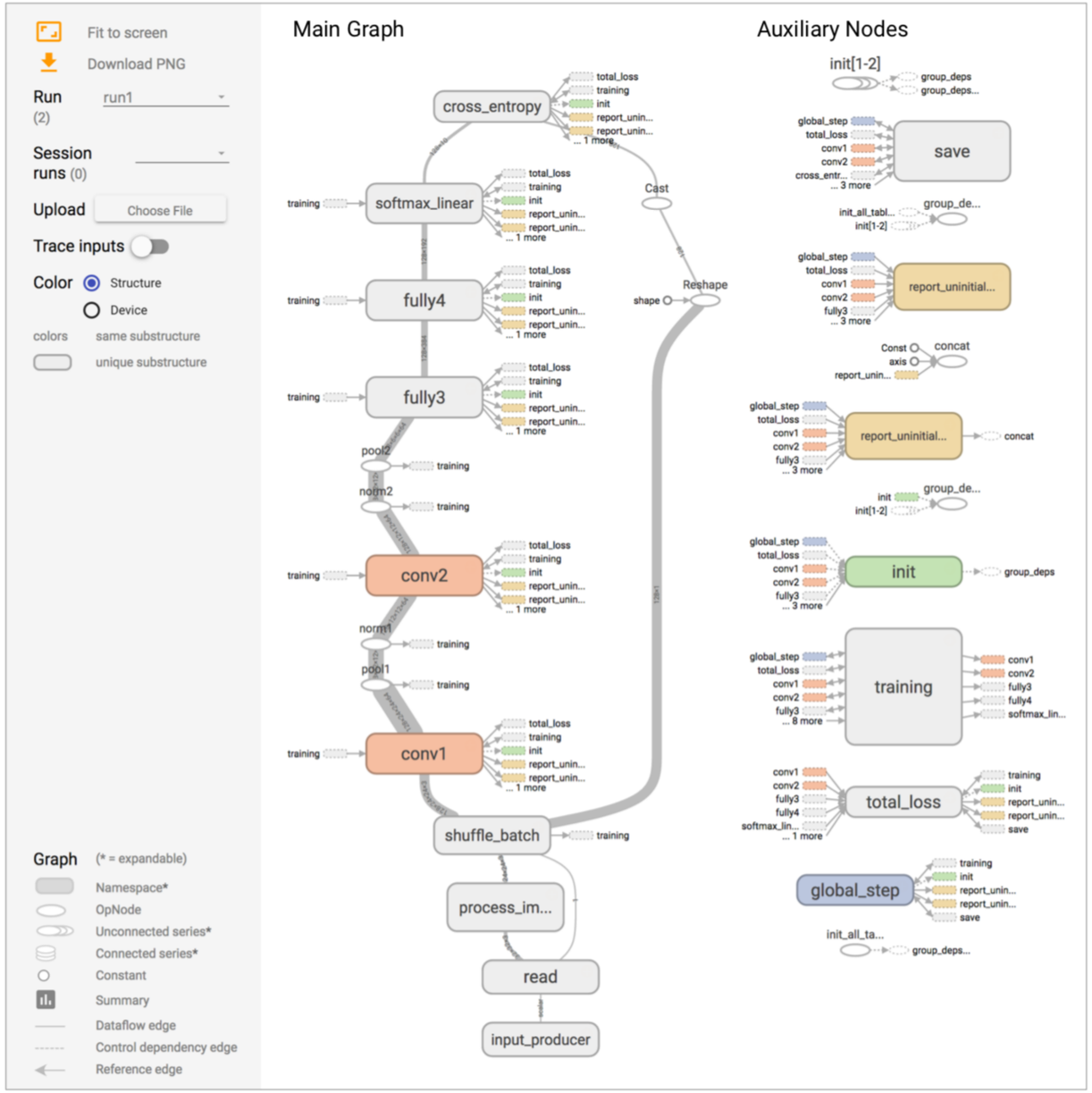
Why?
to ease understanding, explain and convince to reach better adoption of AI

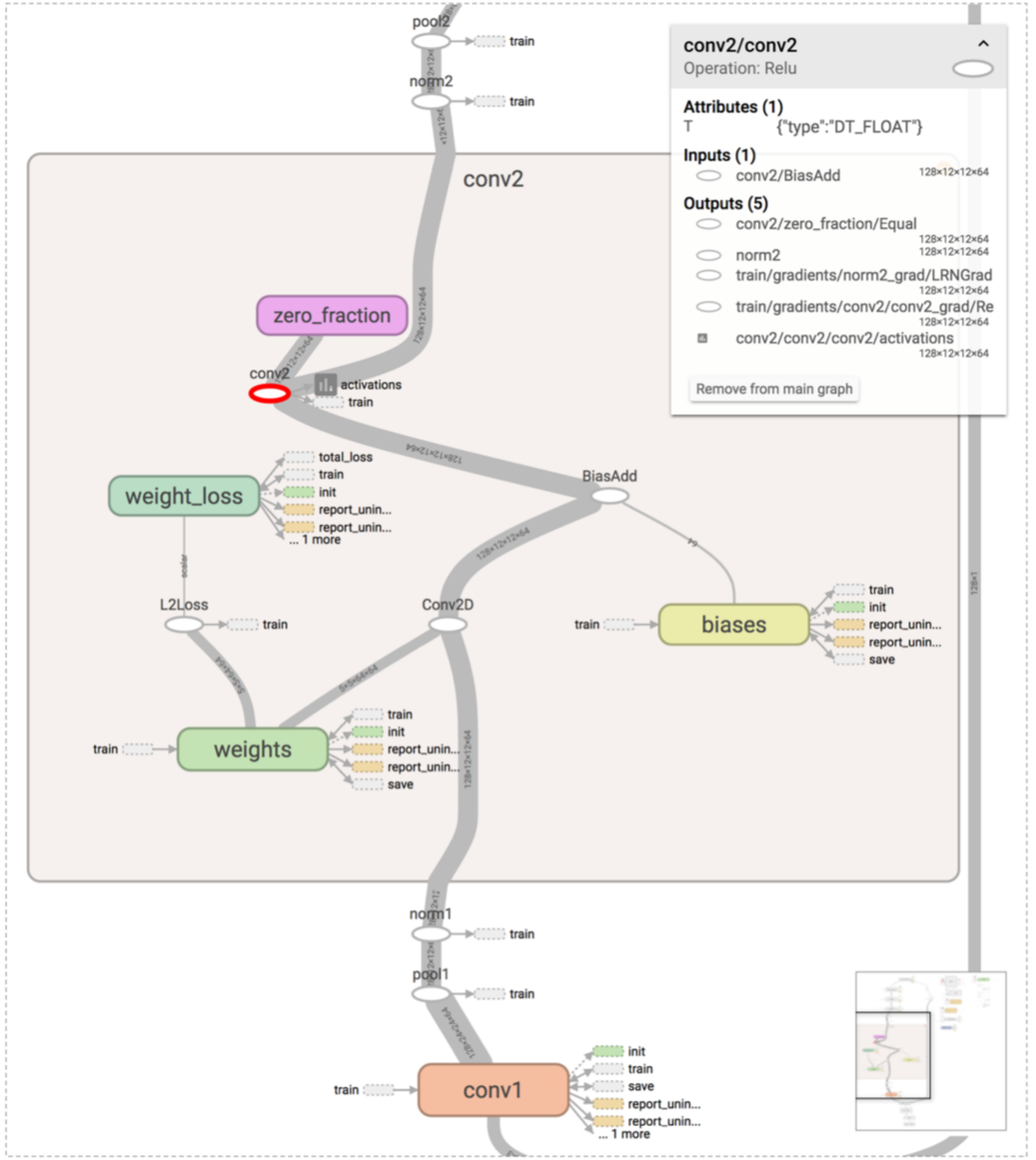
For whom?
General public, citizens, domain experts, AI experts, policy makers

What?
Internal mechanics of a ML model, or a simplified version of it
Relationships between input and output
Reasons why a given output is provided in specific cases

mix

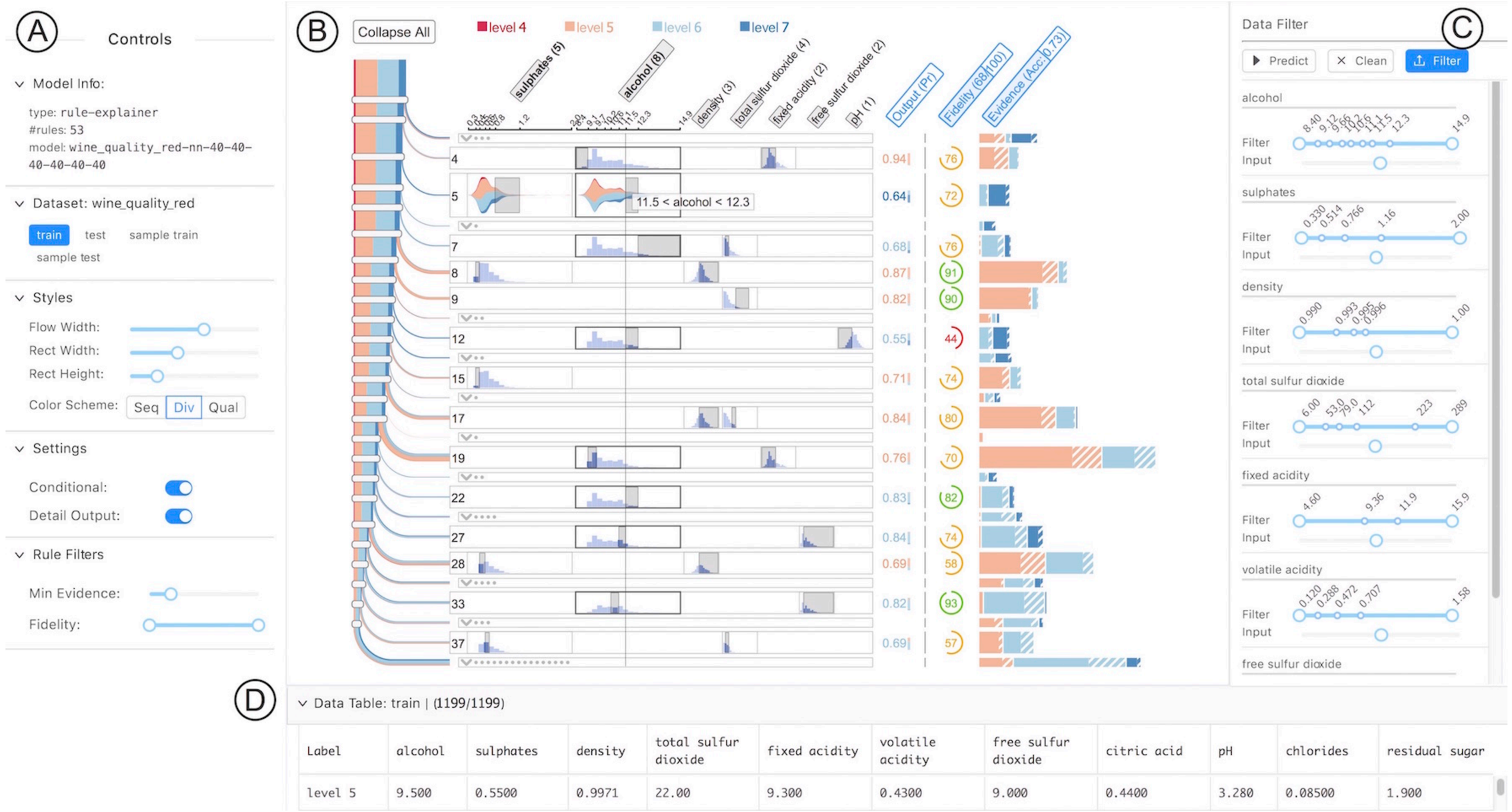
for AI experts





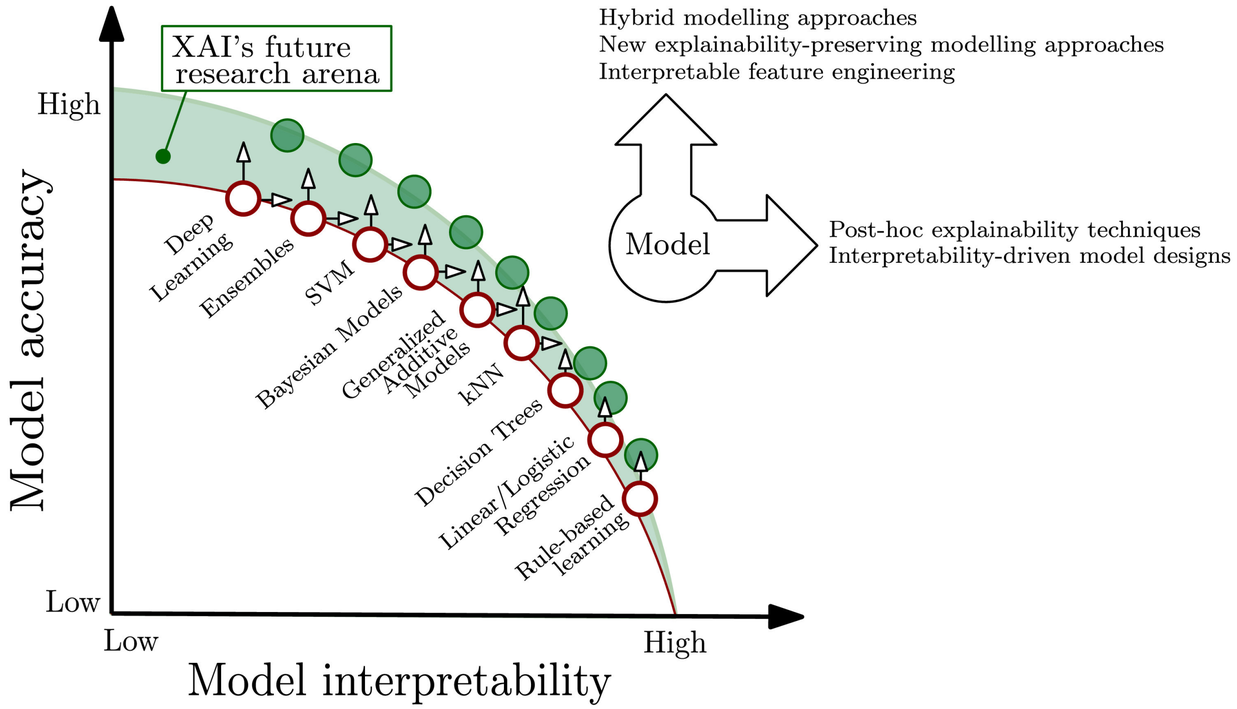
TensorFlow Graph Visualizer, Wongsuphasawat et al. 2018

for domain experts - maybe



RuleMatrix (Ming et al., 2019, https://github.com/rulematrix/rule-matrix-py)

Explainability

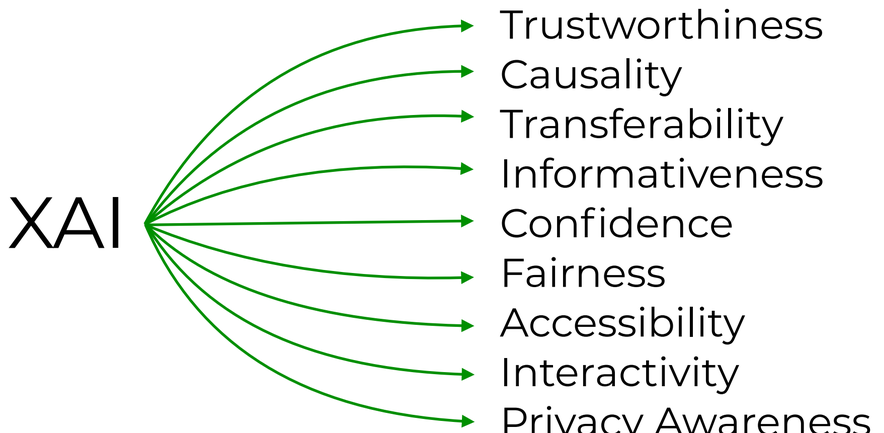


A. Barrido Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Information Fusion 58, 2020

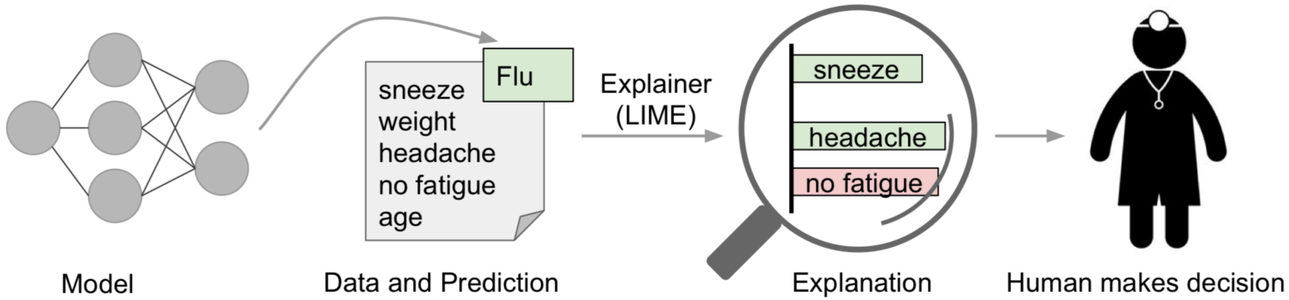
Explainability comes at a price: **accuracy**

How to decide what is more important (transparency/ explainability or accuracy) depends on context:
- credit scoring
- insurances
- cancer detection
- self-driving cars

Possible solution:
Explainable AI (XAI)



LIME



LIME (**L**ocal **I**nterpretable **M**odel-agnostic **E**xplanations): Ribeiro et al., KDD, 2016 (arXiv:1602.04938)

Explains individual predictions of any classification model

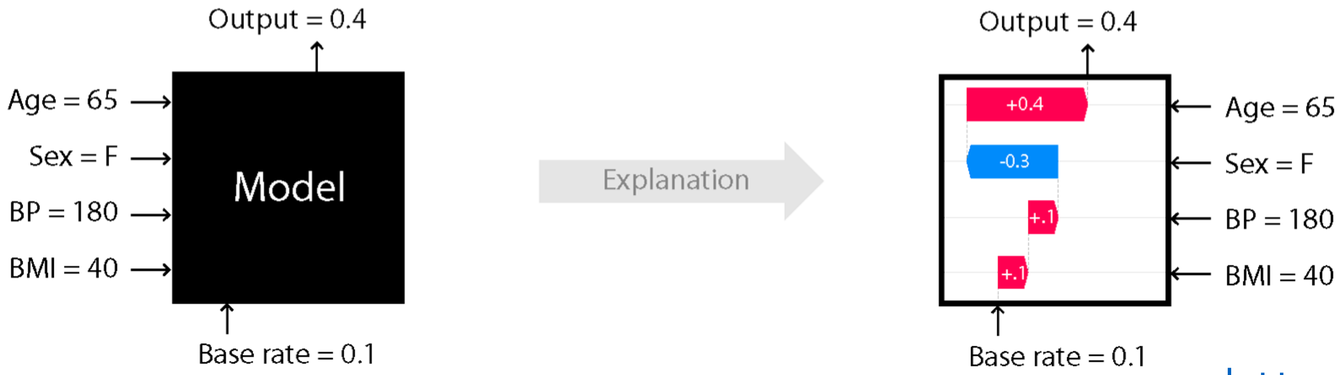
Explanation by simplification: Linear approximation of the model around the vicinity of a particular instance

https://github.com/marcotcr/lime

Shapley values

- Concept in cooperative game theory
- contribution of each individual feature to the model output
- SHAP: average contribution of each feature to each prediction for each sample based on all possible features + different visualisations

SHAP



https://github.com/slundberg/shap

Collection of tools for explainable AI:

https://github.com/EthicalML

Exercise in Explainability

Ethical AI - Explainability Self-Study Exercise.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

Share

Table of contents

- Explainability Self-Study Exercise
  - 1a) Load and inspect the Wine Quality dataset
  - 1b) Modelling
  - 1c) Explanation with LIME
  - 1d) Explanation with SHAP
    - Individual explanations
    - Many observations
      - Run this cell to see a solution
      - Run this cell to see a solution
      - Run this cell to see a solution
      - Run this cell to see a solution
      - Run this cell to see a solution
    - Global interpretability
      - Feature importance
      - Dependence Plots
- 2) Train what you learned above with a new dataset
  - **2b) Explanation with LIME**
- Section

+ Code + Text Copy to Drive RAM Disk Editing

Explainability Self-Study Exercise

Author: Nina Nowak, Senior Data Scientist at Combient Mix

In this notebook we will try to explain the output of a simple classification model using the Python modules LIME and SHAP. You do not need to know how to build machine learning models, but to do the exercise segments you will need some basic knowledge of Python.

To run a cell: press SHIFT+ENTER

Let's start and first install and import some python modules

```
[ ]  !pip install lime
     !pip install shap
```

```
[1]  import pandas as pd
     import sklearn
     import matplotlib.pyplot as plt
     import numpy as np
     import seaborn as sns
     import io
```

1a) Load and inspect the Wine Quality dataset

We use the Wine Quality Dataset https://archive.ics.uci.edu/ml/datasets/wine+quality

Bias and Fairness

References

Solon Barocas & Andrew D. Selbst, “Big Data’s Disparate Impact”, 104 Calif. L. Rev. 671, 2016

Songül Tolan, Fair and Unbiased Algorithmic Decision Making, Digital Economy Working Paper 2018-10; JRC Technical Reports.

Ninareh Mehrabi et al. “A Survey on Bias and Fairness in Machine Learning.” arXiv:1908.09635, 2019

Sahil Verma & Julia Rubin, “Fairness definitions explained”, FairWare ’18: Proc. Int. Workshop on Software Fairness, 2018

Dana Pessach & Erez Shmueli, “Algorithmic Fairness”, arXiv:2001.09784, 2020

https://research.google.com/bigpicture/attacking-discrimination-in-ml/

edX MOOC: EdinburghX: DEI01x, Data Ethics, AI and Responsible Innovation

Solon Barocas, Moritz Hardt & Arvid Narayanan, “Fairness in Machine Learning”, https://fairmlbook.org/

.

Bias definition

Bias
- (orig.) when something is at slant or angle
- (coll.) personal and unreasonable judgement about someone or something
- (tech.) systematic deviation from a true state, e.g. introduced by errors in sampling

Bias creates a false representation of something.

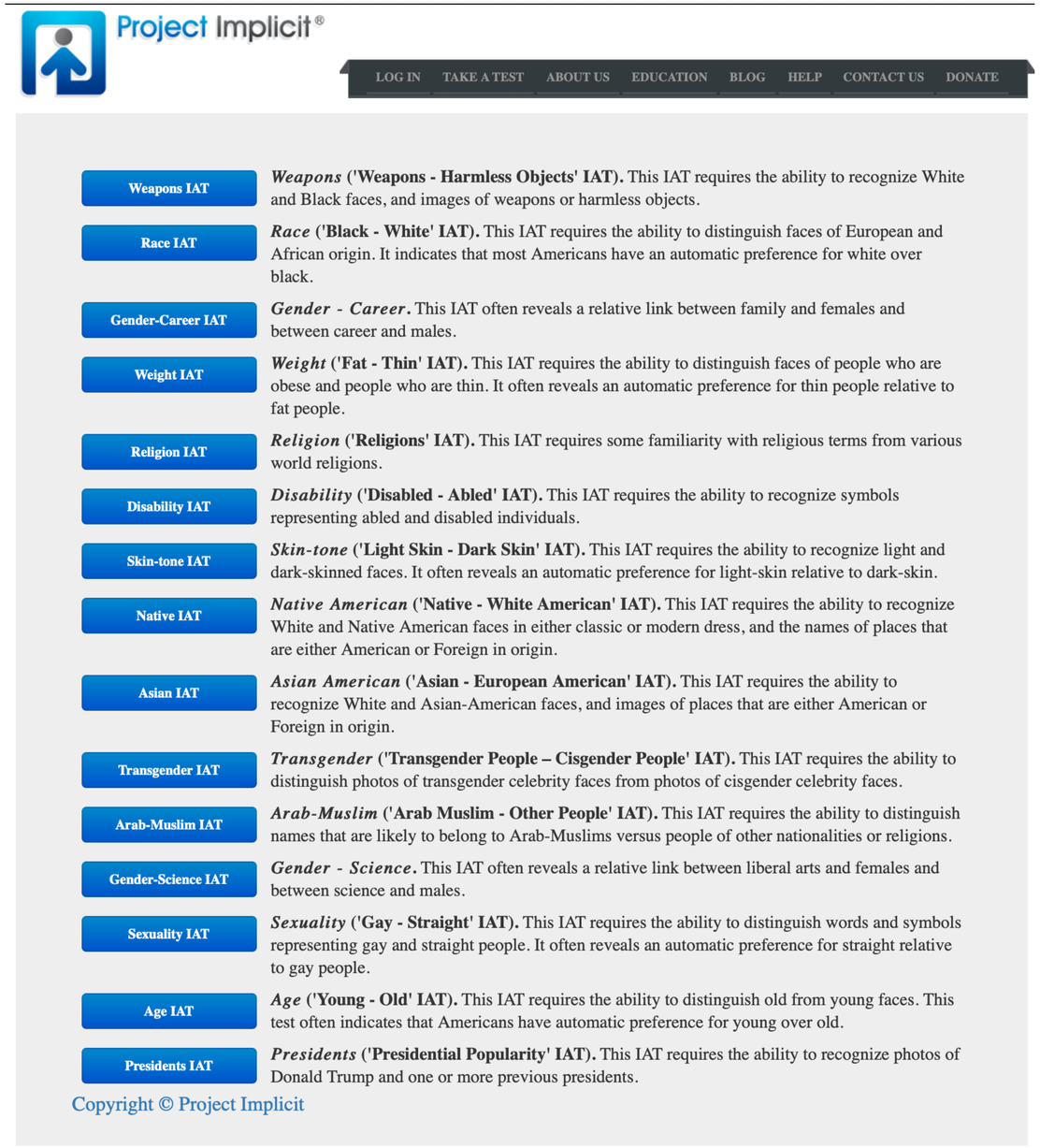
Protected attributes (Article 21 of the Charter of Fundamental Rights of the European Union, **non-discrimination law**):

- Sex
- Race
- Colour
- Ethnic or social origin
- Genetic features
- Language
- Religion or belief
- Political or any other opinion
- Membership of a national minority
- Property
- Birth
- Disability
- Age
- Sexual orientation

Where do biases come from?

Explicit and implicit biases exist in society.

We data scientists need to make sure that we do not propagate these biases further into our models. Models should be developed in a way that every person is treated fairly.



https://implicit.harvard.edu/implicit/

How bias enters the ML model

5 main mechanisms:

1. Defining the ‘target variable’ and ‘class labels’
   - fraud/not fraud
   - credit score or creditworthiness
   - ‘good’ employee
2. Training Data
   - historical bias (e.g. Amazon recruitment tool)
   - sampling bias (e.g. Facial recognition)

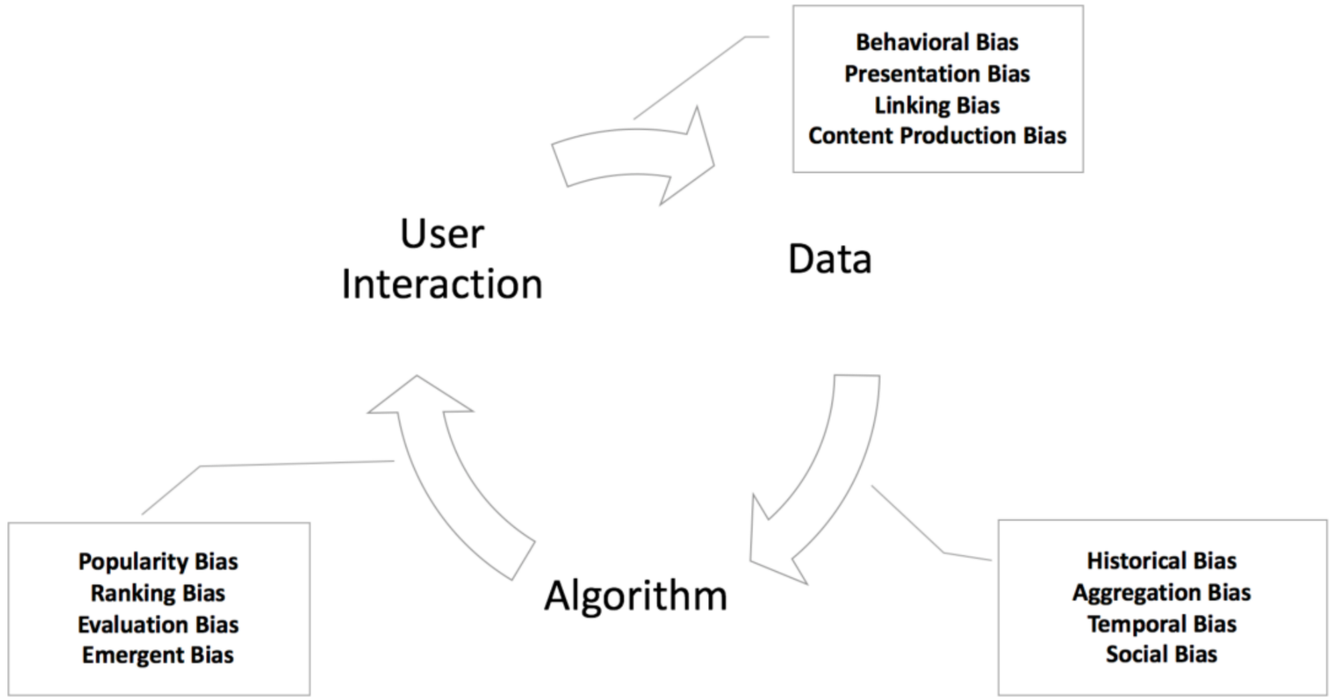
   Often, unbiased training data does not exist
3. Feature selection and encoding
4. Proxies

   Class membership correlates with other features, e.g. ethnicity with zip code, gender with gap in CV, …
   ‘Redundant encodings’
5. Masking

   Intentional discrimination by exploiting 1.-4.

Solon Barocas & Andrew D. Selbst, Big Data's Disparate Impact, 104 CALIF. L. REV. 671 (2016).

Types of Biases in Machine Learning



Mehrabi, Ninareh et al. “A Survey on Bias and Fairness in Machine Learning.” ArXiv abs/1908.09635 (2019)

Mehrabi, et. al. define 23 types of bias relevant to machine learning:

1. **Historical Bias.** Historical bias is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection.
2. **Representation Bias.** Representation bias happens from the way we define and sample from a population.
3. **Measurement Bias.** Measurement bias happens from the way we choose, utilize, and measure a particular feature.
4. **Evaluation Bias.** Evaluation bias happens during model evaluation.
5. **Aggregation Bias.** Aggregation bias happens when false conclusions are drawn for a subgroup based on observing other different subgroups or generally when false assumptions about a population affect the model’s outcome and definition.
6. **Population Bias.** Population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population represented in the dataset or platform from the original target population.
7. **Simpson’s Paradox.** According to Simpson’s paradox, a trend, association, or characteristic observed in underlying subgroups may be quite different from association or characteristic observed when these subgroups are aggregated.
8. **Longitudinal Data Fallacy.** Observational studies often treat cross-sectional data as if it were longitudinal, which may create biases due to Simpson’s paradox.
9. **Sampling Bias.** Sampling bias arises due to non-random sampling of subgroups.
10. **Behavioral Bias.** Behavioral bias arises from different user behavior across platforms, con- texts, or different datasets.
11. **Content Production Bias.** Behavioral bias arises from different user behavior across platforms, con- texts, or different datasets.
12. **Linking Bias.** Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behavior of the users.
13. **Temporal Bias.** Temporal bias arises from differences in populations and behaviors over time.
14. **Popularity Bias.** Items that are more popular tend to be exposed more. However, popularity metrics are subject to manipulation—for example, by fake reviews or social bots.
15. **Algorithmic Bias.** Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm.
16. **User Interaction Bias.** User Interaction bias is a type of bias that can not only be observant on the Web but also get triggered from two sources—the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction.
17. **Social Bias.** Social bias happens when other people’s actions or content coming from them affect our judgment.
18. **Emergent Bias.** Emergent bias happens as a result of use and interaction with real users. This bias arises as a result of change in population, cultural values, or societal knowledge usually some time after the completion of design.
19. **Self-Selection Bias.** Self-selection bias is a subtype of the selection or sampling bias in which subjects of the research select themselves.
20. **Omitted Variable Bias.** Omitted variable bias occurs when one or more important variables are left out of the model.
21. **Cause-Effect Bias.** Cause-effect bias can happen as a result of the fallacy that correlation implies causation.
22. **Observer Bias**. Observer bias happens when researchers subconsciously project their expectations onto the research.
23. **Funding Bias.** Funding bias arises when biased results are reported in order to support or satisfy the funding agency or financial supporter of the research study.

Fairness

Trolley Problem-style tradeoffs when implementing data-based systems: we gain some value at the cost of some other value. Often the cost is in fairness because of biases in the algorithms.

Legal definitions of fairness (anti-discrimination law):

- Avoid direct discrimination (‘*disparate treatment*’): person is treated differently based on membership to a protected group
- Avoid indirect discrimination (‘*disparate impact*’): apparently neutral rule leads to outcomes that differ based on membership to a protected group

There are **20+ statistical definitions of fairness**, which are incompatible with each other, except in the most trivial cases.

Fairness definitions

Individual fairness
Give similar predictions to similar individuals

Fairness through unawareness
An algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process
- Usually results in indirect discrimination via proxies
- Not even collecting sensitive data makes it impossible to discover/measure (un)fairness

Fairness through awareness
Similar individuals should have similar classification. Similarity is defined via a distance metric
- in practice hard to determine an appropriate distance metric

Counterfactual fairness
Based on causal reasoning. A decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group
- in practice complicated to use and would suffer significant loss in accuracy when all proxy features are eliminated

Fairness definitions

Group/Statistical fairness
Treat different groups equally

3 types of group fairness, based on:

A. predicted outcome
B. combination of predicted and actual outcomes
C. combination of predicted probabilities and actual outcomes

Most statistical measures of fairness rely on these metrics:

Table 1: Confusion Matrix

		Predicted Classification $\hat{Y}=1$	Predicted Classification $\hat{Y}=0$	
Outcome	$Y=1$	True Positives (TP)	False Negatives (FN)	False Negative Rate (FNR) $FN/(TP+FN)$
Outcome	$Y=0$	False Positives (FP)	True Negatives (TN)	False Positive Rate (FPR) $FP/(FP+TN)$
		False Omission Rate (FOR) $FP/(TP+FP)$	False Discovery Rate (FDR) $FN/(FN+TN)$	

Cross-tabulation of actual and predicted outcomes.

Songül Tolan, Fair and Unbiased Algorithmic Decision Making, Digital Economy Working Paper 2018-10; JRC Technical Reports.

A. Based on predicted outcome
Demographic Parity

The probability of every outcome should be the same, or in practice at least similar, for each group.

$$\mathbb{E}[\hat{Y} = 1 \mid A = a] = \mathbb{E}[\hat{Y} = 1 \mid A = b]$$

Customer	Income	credit score	Gender	Repayment
A	100000	1	F	0
B	200000	2	F	1
C	300000	3	M	1
D	300000	4	F	0
E	400000	4	M	1
F	500000	3	F	1
G	600000	5	M	0
H	600000	5	M	1

The share of positive and negative outcomes should be the same for each group:

$$P = \frac{TP + FP}{G} \qquad N = \frac{TN + FN}{G}$$

G = group size

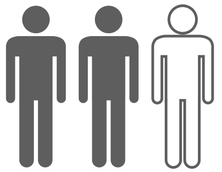
Songül Tolan, Fair and Unbiased Algorithmic Decision Making, Digital Economy Working Paper 2018-10; JRC Technical Reports.
edX MOOC: EdinburghX: DEI01x, Data Ethics, AI and Responsible Innovation

A. Based on predicted outcome
Demographic Parity

Classifier: Customers with credit score >3 will repay loan

Customer	Income	credit score	Gender	Repayment
A	100000	1	F	0
B	200000	2	F	1
C	300000	3	M	1
D	300000	4	F	0
E	400000	4	M	1
F	500000	3	F	1
G	600000	5	M	0
H	600000	5	M	1

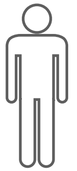
Positives:



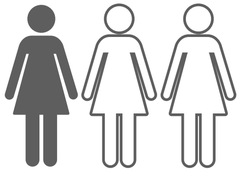
75%

25%

Negatives:



25%



75%

Four-fifth rule: If the selection rate for a certain group is <80% of that of the group with the highest selection rate, there is adverse impact on that group

Songül Tolan, Fair and Unbiased Algorithmic Decision Making, Digital Economy Working Paper 2018-10; JRC Technical Reports.
edX MOOC: EdinburghX: DEI01x, Data Ethics, AI and Responsible Innovation

B. Based on a combination of predicted and actual outcomes
Error rate balance

If the classifier gets it wrong, it should be equally wrong for all groups.

$$\mathbb{E}[\hat{Y} = 1 \mid Y = 0, A = a] = \mathbb{E}[\hat{Y} = 1 \mid Y = 0, A = b]$$ predictive equality

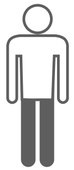
$$\mathbb{E}[\hat{Y} = 0 \mid Y = 1, A = a] = \mathbb{E}[\hat{Y} = 0 \mid Y = 1, A = b]$$ equal opportunity

FPR:

1

0.5

FNR:



0.33

1

The false positive rate (FPR) ***or*** the false negative rate (FNR) should be the same for each group:

$$FPR = \frac{FP}{FP + TN} \qquad FNR = \frac{FN}{TP + FN}$$

Classifier: Customers with credit score >3 will repay loan

Songül Tolan, Fair and Unbiased Algorithmic Decision Making, Digital Economy Working Paper 2018-10; JRC Technical Reports.
edX MOOC: EdinburghX: DEI01x, Data Ethics, AI and Responsible Innovation

B. Based on a combination of predicted and actual outcomes
Equalised odds

The probability of a person in the positive class being correctly assigned a positive outcome and the probability of a person in a negative class being incorrectly assigned a positive outcome should be the same for all groups

TPR:

0.66

0

FPR:

1

0.5

The true positive rate (TPR) ***and*** the false positive rate (FPR) should be the same for each group:

$$TPR = \frac{TP}{TP + FN} \qquad FPR = \frac{FP}{FP + TN}$$

Classifier: Customers with credit score >3 will repay loan

Songül Tolan, Fair and Unbiased Algorithmic Decision Making, Digital Economy Working Paper 2018-10; JRC Technical Reports.
edX MOOC: EdinburghX: DEI01x, Data Ethics, AI and Responsible Innovation

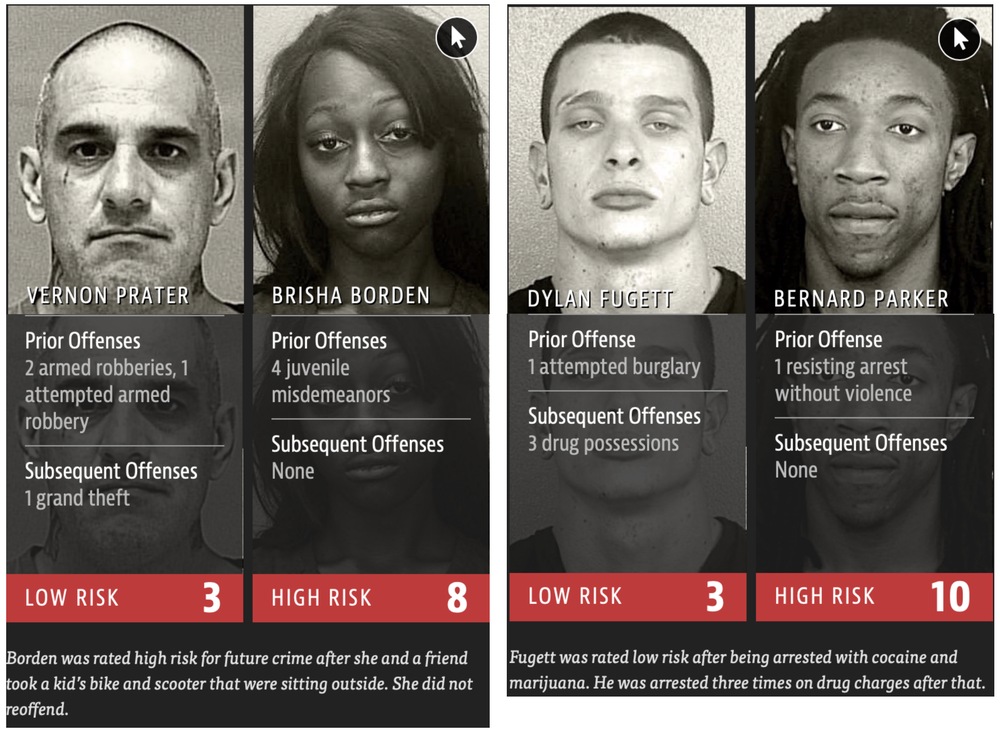
C. Based on a predicted risk scores and actual outcomes
Calibration

For a given risk score, the proportion of people who are considered actually at risk is the same across protected groups.

$$\mathbb{E}[Y = 1 \mid R = r, A = a] = \mathbb{E}[Y = 1 \mid R = r, A = b), \forall r \in R]$$

Example: Loan applicants with an estimated average 10% probability of default. Calibration requires that men and women default at similar rates.

Limitations of fairness measures



COMPAS
A decision support tool used in the US to assess the likelihood of a defendant to become a recidivist.
Trained on questionnaires filled by defendants (137 questions, no question about race).
Goal: reduce societal risk

ProPublica: NGO, investigative journalism
They found that COMPAS was biased against blacks.
Whites were more likely to reoffend, whereas blacks were less likely to reoffend than predicted.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

Northpointe said their algorithm is fair because it is **calibrated**, i.e. for the same risk score, a comparable % of black defendants reoffend in comparison to white defendants.
ProPublica used **error rates** to assess the algorithm.

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Example: COMPAS

Blacks		Predicted COMPAS score Low	High
actual recidivism	0	990	805
	1	532	1369

Whites		Predicted COMPAS score Low	High
actual recidivism	0	1139	349
	1	461	505

	Calibration	Error rate	
Blacks	$\frac{TP}{TP+FP} = 0.630$	$FPR = \frac{FP}{FP+TN} = 0.448$	$FNR = \frac{FN}{FN+TP} = 0.280$
Whites	$\frac{TP}{TP+FP} = 0.591$	$FPR = \frac{FP}{FP+TN} = 0.235$	$FNR = \frac{FN}{FN+TP} = 0.477$
Difference	0.039	0.213	-0.197

Limitations of fairness measures

It is impossible to satisfy more than one fairness criterium simultaneously!
Since the COMPAS algorithm is calibrated, its error rates cannot be equalised.

Conclusion:
Statistical fairness criteria on their own cannot be used as a proof of fairness. They can only provide a starting point for thinking about fairness issues. We have to make choices with what we mean by fairness.

How to mitigate bias

3 categories of algorithms:

1. Pre-processing techniques: remove information correlated to the sensitive attribute. Can only be used to optimise demographic parity and individual fairness
2. In-processing techniques: modify the learning algorithm by changing the objective function or adding a constraint or regularisation. Can be used for any fairness definition
3. Post-processing techniques: labels initially assigned by the model get reassigned based on some function. Can be used for most fairness definitions

Collection of resources and packages:
https://github.com/datamllab/awesome-fairness-in-ai
https://github.com/EthicalML/awesome-production-machine-learning

Tools to measure and mitigate bias

IBM's *AI Fairness 360*

Optimized Pre-processing

Use to mitigate bias in training data. Modifies training data features and labels.

Reweighing

Use to mitgate bias in training data. Modifies the weights of different training examples.

Adversarial Debiasing

Use to mitigate bias in classifiers. Uses adversarial techniques to maximize accuracy and reduce evidence of protected attributes in predictions.

Reject Option Classification

Use to mitigate bias in predictions. Changes predictions from a classifier to make them fairer.

Disparate Impact Remover

Use to mitigate bias in training data. Edits feature values to improve group fairness.

Learning Fair Representations

Use to mitigate bias in training data. Learns fair representations by obfuscating information about protected attributes.

Prejudice Remover

Use to mitigate bias in classifiers. Adds a discrimination-aware regularization term to the learning objective.

Calibrated Equalized Odds Post-processing

Use to mitigate bias in predictions. Optimizes over calibrated classifier score outputs that lead to fair output labels.

Equalized Odds Post-processing

Use to mitigate bias in predictions. Modifies the predicted labels using an optimization scheme to make predictions fairer.

Meta Fair Classifier

Use to mitigate bias in classifier. Meta algorithm that takes the fairness metric as part of the input and returns a classifier optimized for that metric.

Tools to measure and mitigate bias

Fairlearn

User Guide API Docs Contribute **GitHub**

Think fairness. Build for everyone.

A toolkit to assess and improve the fairness of machine learning models.

Assess Mitigate

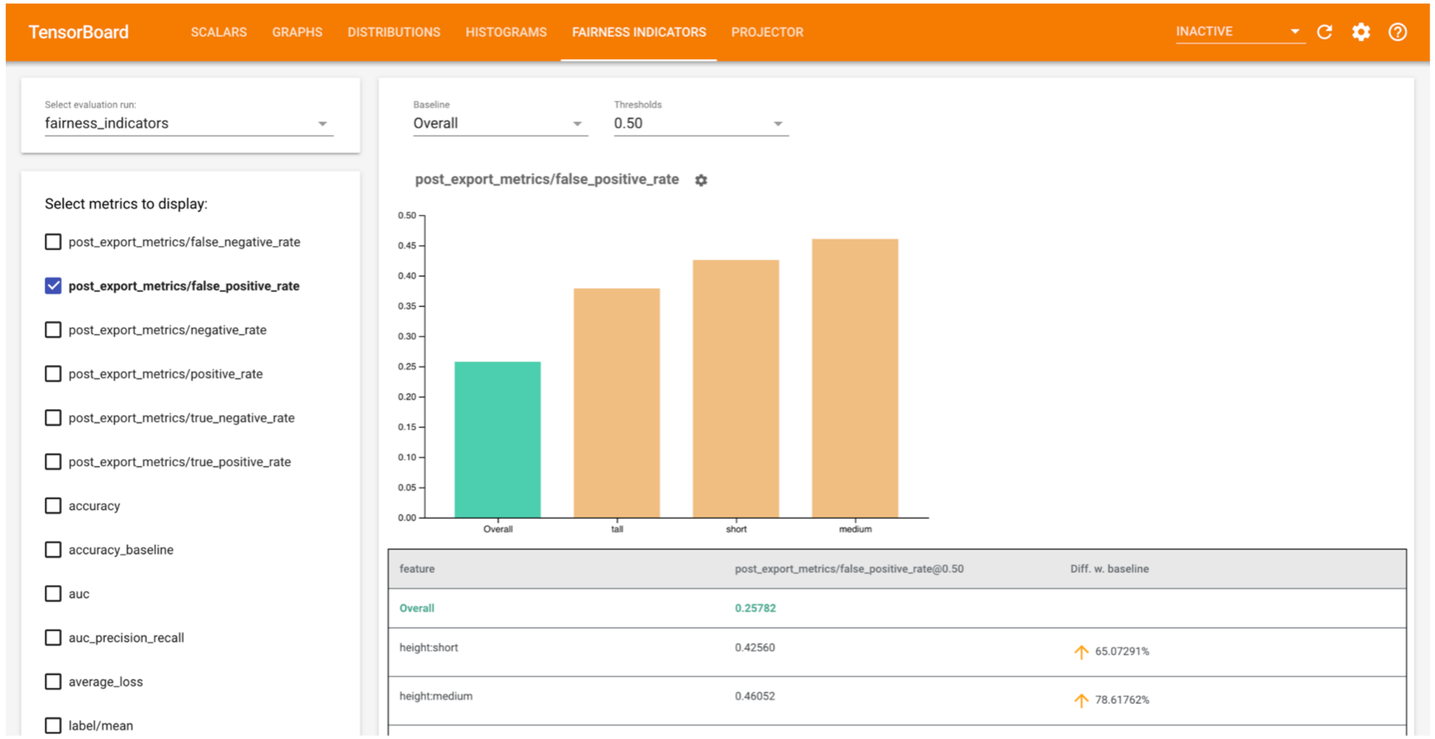
Use common **fairness metrics** and an **interactive dashboard** to assess which groups of people may be negatively impacted.

Get Started **API Docs**

Tools to measure and mitigate bias

TensorFlow > Resources > TensorBoard > Guide

Evaluating Models with the Fairness Indicators Dashboard [Beta]



Fairness Indicators for TensorBoard enables easy computation of commonly-identified fairness metrics for *binary* and *multiclass* classifiers. With the plugin, you can visualize fairness evaluations for your runs and easily compare performance across groups.

https://www.tensorflow.org/tensorboard/fairness-indicators

Tools to measure and mitigate bias

fairmodels part of the DrWhy.AI developed by the MI^2 DataLab 0.1.1

Reference Articles Changelog

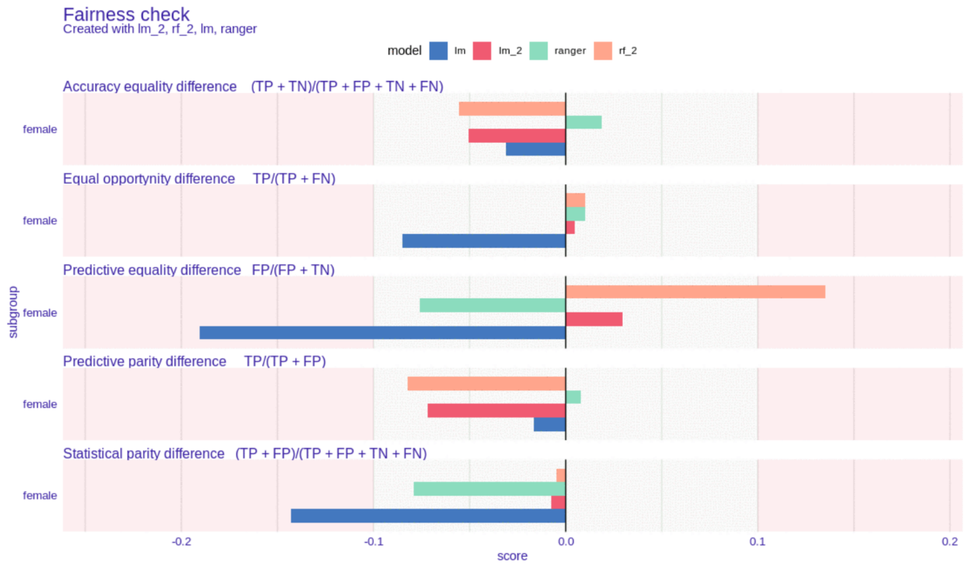
fairmodels

codecov 85% R-CMD-check failing CRAN 0.1.1 DrWhy eXtrAI

Overview

Flexible tool for bias detection, visualization, and mitigation. Uses models explained with DALEX and calculates fairness metrics based on confusion matrix for protected group. Allows to compare and gain information about various machine learning models. Mitigate bias with various pre-processing and post-processing techniques. *Make sure your models are classifying protected groups similarly.*

Preview



Links

Download from CRAN at https://cloud.r-project.org/package=fairmodels

Browse source code at https://github.com/ModelOriented/fairmodels/

Report a bug at https://github.com/ModelOriented/fairmodels/issues

License

GPL-3

Developers

Jakub Wiśniewski
Author, maintainer

Przemysław Biecek
Author

Ethical AI - Fairness Self-Study Exercise.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

Share N

Table of contents

- Ethical AI - Fairness self-study exercise
  - Load and inspect the Adult Income dataset
  - Modelling
  - Measure Biases - Fairlearn Dashboard
  - Measure Biases - Statistical Fairness measures
    - Demographic Parity
    - Error rate balance
    - Equalized Odds
  - Bias mitigation

Section

+ Code + Text Copy to Drive Connect Editing

Ethical AI - Fairness self-study exercise

In this notebook we will try to measure and mitigate fairness of a simple classification model using the Python module Fairlearn. You do not need to know how to build machine learning models, but to do the exercise segments you will need some basic knowledge of Python.

To run a cell: press SHIFT+ENTER

Let's start and first install and import basic python modules

```
!pip install fairlearn
!pip install ipywidgets
```

```
[ ] import pandas as pd
    import sklearn
    import matplotlib.pyplot as plt
    import numpy as np
    import seaborn as sns
```

1. Load and inspect the Adult Income dataset

The Adult Income dataset is from the 1994 United States Census Bureau. The task is to predict whether a given individual makes more than $50,000 a year based attributes such as education, hours of work per week, etc.

```
[ ] # We use the Adult Income Dataset http://archive.ics.uci.edu/ml/datasets/Adult
    # Show the first few rows of the dataset
    adult = pd.read_csv('https://github.com/nillepu/EthicalAI/raw/main/adult-all.csv',sep=',')
    adult.head()
```

https://colab.research.google.com/drive/1uZ0AC3Uls1tgHb5HfCYu0mzMgx3Pjv0Q?usp=sharing

Distributive fairness

How to make fair algorithms that compute allocations of benefit among different people, when resources are limited and people have different preferences?

Theoretical framework: Economics and game theory.

Assumptions: individuals have different preferences and are assumed to be self-interested. Conflict of interest due to limited resources.

Ethical framework: Utilitarianism

Key metrics: Social welfare, equity, maximin, Pareto efficiency

Examples: Housing, loans, ad space, ranking on a hotel recommendation page, computing resources, ...

Distributive fairness metrics

$W = W(U_1, \ldots, U_i, \ldots, U_N)$ W = Welfare, U_i = utility/well-being of individual i

Utilitarian Social Welfare: maximise the sum of utilities of all agents
does not take variations in individual utilities into account, so can lead to small number of extremely happy people at cost of many unhappy

$$W = \sum_{i=1}^{N} U_i$$

Equity: minimise the differences between the outcomes for all individuals
(e.g. sum over pairwise distance)

$$W = \min_{i \neq j} \sum_{i,j=1}^{N} d(U_i, U_j)$$

Maximum Nash Welfare: maximise the product of utilities used in bargaining situations
example: when 2 people split a cake, MNW would be maximal if both get 0.5

$$W = \prod_{i=1}^{N} U_i$$

Distributive fairness metrics

Egalitarian Social Welfare: *(Maximin)*
maximise the outcome for the participant who is worst off
does not guarantee how efficient outcome is for the society as a whole

Pareto efficiency:
allocation where no one's outcome can be improved without making other people worse off
requires pair-wise comparison of all outcomes
unclear which to choose when there are >1 Pareto-optimal outcomes

Envy-freeness:
guarantees that everyone feels what they get is at least as good as what others get
hard to achieve in reality, so often side payment to balance benefit

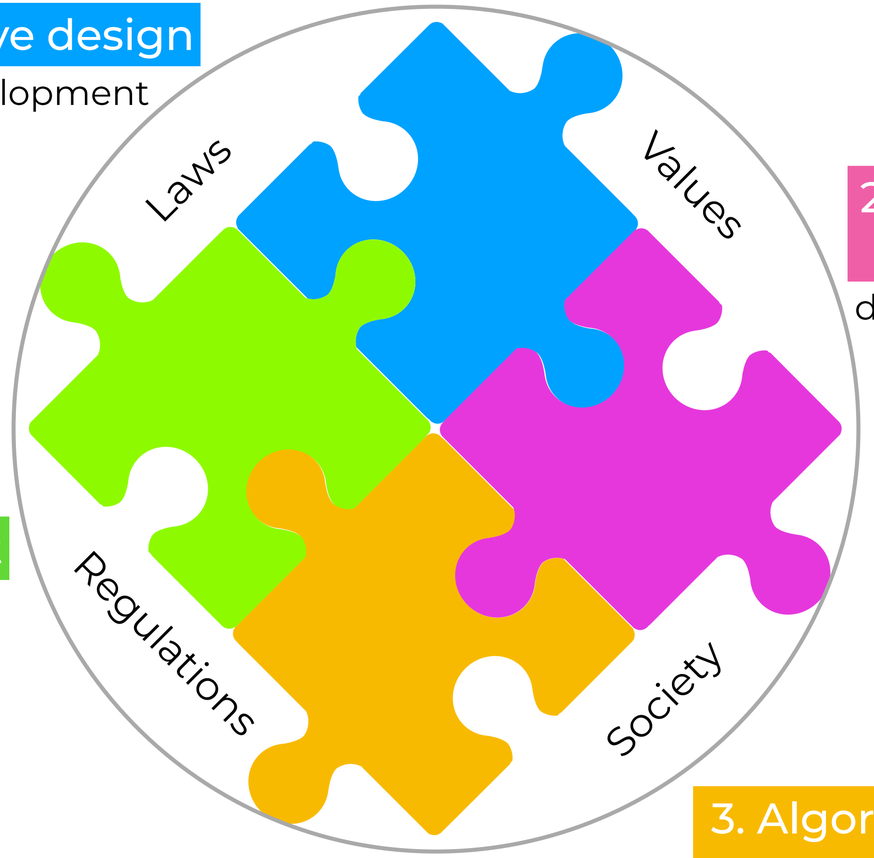
Problems:

- How to measure utility/people's values or preferences in algorithmic systems
- People do not behave rationally and entirely self-interested

How to create responsible AI?

1. Value-sensitive design

before/during development



2. Statistical/ML/technical methods

during/after development

4. Algorithmic audit

after deployment

3. Algorithmic impact assessment

before/after deployment