Ethics and Accountability of Algorithmic Decision Making Systems

Prof. Dr. K.A. Zweig
TU Kaiserslautern
Algorithm Accountability Lab
@nettwerkerin
AI raises fears

AI will create poetry...

... and penalty
Algorithmic Decision Making Systems (ADM)

Algorithm

SCHUFA (credit worthiness)

Car insurance

Klasse 1

Klasse 2

Klasse 3

Scoring or Classification
Who should judge humans?
Are humans the best option to judge other humans?
Mankind – so irrational!

- Study: Judges have to review prison release proposals regularly.
- Shown: Time from last break reduced likelihood for a positive decision\(^1\).
- Many more studies seem to show:
  - Humans are irrational and biased.

---
Problematic situation in the USA

- Second highest incarceration rate worldwide.

- 6x higher rate of Afroamericans and 2x more of Latinos.

- Dramatic prognosis: every third boy at the age of 10 now will be in prison at least once in his life.
• American civil liberties states:
• ADM systems need to be used in all stages of the legal process, ...
• ... to ensure fairness and objectivity.
• They propose that computer should learn the necessary decision rules from data.
Can computers learn?
What is learning?

Simple:
Repeat some learned behavior in some defined situation.

Generalized:
Choose the correct behavior from a range of possibilities in the same kind of situation.
Sebastian learns „hot“ and „warm“

Too cautious, No steam, please
eats only cold meals

Too daring
Sebastian learns...

• By **feedback**: unexpectedly hot, unexpectedly cold

• By **saving rules in some structure**: in neurons and their connections.

• By **many data points** (experiences).

• By **generalizing the learned rules**.
Computers learn...

By giving them **a structure** for saving learned rules.

By giving them **feedback**.

By learning.
“Learn” from correlations
You have to welcome a new employee. Is it Mr. or Ms Miller?

You know that the person gets less than $25/h. Is it rather Mr. or Ms Miller?
“Learning” with SVMs
Aggressive criminals

Innocent citizens
What do you think of Ms Miller?

5.5 Sanftosan
4.0 Kriminolin
Aggressive criminals

Innocent citizens
Aggressive criminals

Innocent citizens

?
“It is better that ten guilty persons escape than that one innocent suffer.”

William Blackstone, Rechtsphilosoph, 1760

"I am more concerned with bad guys who got out and released than I am with a few that, in fact, were innocent."

Dick Cheney, ehemaliger Vizepräsident der USA,
Terrorists

Innocent citizens

Graph showing the relationship between Kriminolin and Sanftosan.
Unidentified financial fraud

Actually innocent

Data quality

Sanftosan
Learning with formulae

Recidivism risk assessment of criminals
Data

• Data Mining methods use, e.g.:
  • Age at first arrest
  • Age now
  • Financial situation
  • Criminal relatives (!)
  • Gender
  • Number and kind of previous convictions
  • Time point of last criminal action
  • A survey
  • But (of course) not the race of a person.

• To learn something, we need this data plus the information whether the person has recidivated or not.
Approaches: Regressions

• In practice, algorithm designer very often decide which data most likely correlate with „recidivism“.

• The result of the algorithm should be a single number.

• The higher the number, the higher the risk.

• Example formula:

\[
3 \times \text{# previous convictions} \\
- 2 \times \text{# days since last arrest} \\
+ 3 \times (1 \text{ if man, 0 else}) \\
+ 2.5 \times (1 \text{ if violent act involved, 0 else}) + ...
\]
In general

\[ w_1 \times \# \text{ previous convictions} \]
\[ - w_2 \times \# \text{ days since last arrest} \]
\[ + w_3 \times (1 \text{ if man}, 0 \text{ else}) \]
\[ + w_4 \times (1 \text{ if violent act involved}, 0 \text{ else}) + \ldots \]

The computer determines the weights and gets a feedback on its predictions and the actual observation of recidivism in that individual.
Quality of an algorithm
„Learning“ of weights

- Algorithm 'tries out' weights and computes resulting risk for all persons in a test data set.
- Evaluates how many of the real recidivists get high risk scores.
- The weighting that maximizes this will be used for all further predictions.

Green balls: non-recidivating criminals; Red balls: recidivating criminals.

Optimal sorting: all reds on top, all greens on the bottom.

Quality measure: pairs of red/green balls where red is on top of the green.
Oregon Recidivism Rate Algorithm

• 72 of 100 pairs are correctly sorted (72% success rate! Yeah!)
• Does this resemble they way judges make a decision?
• No, instead of judging pairs, they see a sequence of defendants, of which they are most interested in the ones with highest risk.
• Experience guides where to cut the risk score:
  • E.g., recidivism rate of young criminals is about 20%.
Optimal Sorting

Expected 20% recidivists
Possible sorting of an algorithm with 75% correctly sorted pairs

Expected 20% recidivists
Buying a hunting dog, to shepherd!

Buying that software is like...
From scoring to classification

• ACLU states: criminals should be sorted into three categories...

• ... with low, medium and high risk, respectively.

10/24 = 42%

9/29 = 31%

2/37 = 5%
Statistical predictions of human behavior
Weather forecasts
40% a criminal....

• If humans were cats with 7 lives, to have a 40% risk of recidivism it would mean that they expectedly recidivate in three of them.

• But no! Humans are no cats.

• Algorithmically legitimated prejudices
  • Of 100 persons that are „like this person“ 40% are recidivating.
Rule of thumb

AI is used primarily where there are no simple rules!

They often search for patterns in highly noisy data.

The patterns are of a statistical nature.

Often try to identify a small group of people (Problem of imbalance)
Can algorithms discriminate people?
Equality

This is what happens when you search for „CEO“ on Google
And this, if I search for ‘boss’ on Pixabay....
Discrimination

- Google shows job ads with a lower average salary to female surfers.
- Based on one perspective, recidivism risk assessment is rassistic.
- Discrimination in training data will be learned.
- If training data contains too little data about minorities, their properties will not be learned.
Algorithms in a democracy
In general

In principle, ADM systems can be used for many different, difficult questions:

- Automatic performance evaluation
- Credit approval
- Job application evaluation.
- Performance evaluation of employees.
- Algorithms that predict the time point of death (real!)
- Terrorist identification
- ...
Your task today....

Develop an ADM system that identifies terroristic couriers!
Design process

Data Scientist

Researcher

Development of analytic method → Implementation

Method selection

Development of analytic method → Implementation

Development of analytic method → Implementation

Person or Institution

Operationalization

Data collection

Data collection

Data selection

Person or Institution

Data

Feedback

Interpretation of result

Decision of action
Capturing terrorists with network analysis

From GSM metadata, we can measure aspects of each selector's pattern-of-life, social network, and travel behavior.
Terrorist identification SKYNET

We’ve been experimenting with several error metrics on both small and large test sets

<table>
<thead>
<tr>
<th>Training Data</th>
<th>Classifier</th>
<th>Features</th>
<th>100K Test Selectors</th>
<th>55M Test Selectors</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>False Alarm Rate at 50% Miss Rate</td>
<td>Mean Reciprocal Rank</td>
</tr>
<tr>
<td>None</td>
<td>Random</td>
<td>None</td>
<td>50%</td>
<td>1/23k (simulated)</td>
</tr>
<tr>
<td>Known Couriers</td>
<td>Centroid</td>
<td>All</td>
<td>43%</td>
<td>1/27</td>
</tr>
<tr>
<td>+ Anchor Selectors</td>
<td>Random Forest</td>
<td>Outgoing</td>
<td>0.008%</td>
<td>1/14</td>
</tr>
</tbody>
</table>

Random Forest trained on Known Couriers + Anchor Selectors:
- 0.008% false alarm rate at 50% miss rate
- 46x improvement over random performance when evaluating its tasked precision at 100

These are 4,400 innocent persons to catch 50% of the (suspected) terrorists.

https://theintercept.com/document/2015/05/08/skynet-courier/
Most suspicious person according to algorithm is...
How good are these robo-judges?

• Very bad: COMPAS
  • High risk category:
    • General recidivism: correct in 50% of all individuals!
    • Serious crimes: only 20% correct!

• An American terrorist identification system boasts:
  • „Only 0.008% false positives!“
  • With 55 million inhabitants these are about 4,400 innocents to identify a few hundred.

• However, in detecting cancer they are sometimes better than physicians.
Socio-informatic system analysis
Problems with the embedding of an ADM into a social process

• Deciders follow an ‘attention economy’.
• „Best practice“ might require the usage of the software.
• **Delegation of responsibility!**
• Sometimes a false-negatively categorized person cannot prove the decision wrong!
  • E.g., rejected applicants for a job,
  • Rejected credits,
  • Suspects kidnapped and kept in camps.
My stance

• ADM systems **could** help to make better decisions.
  • They can search through huge data sets.
  • They could identify new “patterns”.
  • Could avoid discriminations.

• **However**, today, they’re not yet there. And they might not be able to, especially when very few persons have to be identified in a crowd.
Problems of ADM systems in people and risk assessment

1. Who decides what a „good“ decision is?
2. ADM systems result in probabilities, not truths.
3. ADM systems can discriminate
4. The help to identify small groups but with many false positives.
5. ADM systems can change social processes.
6. The reaction of the social system can increase the problem.

@nettwerkerin
Prof. KA Zweig
TU Kaiserslautern
ADM systems to be regulated

... with a learning component,

which directly affects humans.
Do all of them need to be regulated the same?

1. Potential for damage in case of errors

\[ \sum \text{Potential for harm for individual (in case of error)} + \text{Potential for harm for society (in case of errors)} \]

2. Number of competitors and ease of re-evaluation by other ADM systems
Product recommendation

Scoring of objects

Low damage in case of errors

Bit market, easy change

Credit scoring

Monopoly

Potentially high damage in case of errors.

Sort and filter applicants

Employee performance evaluation

Facebook Newsfeed

Automatic tax re-imbursement

Scoring of objects

China Citizen Score

Lethal Autonomous Weapons

Terrorist identification

Employee performance evaluation

Automatic tax re-imbursement

China Citizen Score

Lethal Autonomous Weapons

Terrorist identification

Scoring of objects

Employee performance evaluation

Automatic tax re-imbursement

China Citizen Score

Lethal Autonomous Weapons

Terrorist identification

Scoring of objects

Employee performance evaluation

Automatic tax re-imbursement

China Citizen Score

Lethal Autonomous Weapons

Terrorist identification

Scoring of objects

Employee performance evaluation

Automatic tax re-imbursement

China Citizen Score

Lethal Autonomous Weapons

Terrorist identification

Scoring of objects

Employee performance evaluation

Automatic tax re-imbursement

China Citizen Score

Lethal Autonomous Weapons

Terrorist identification
Low damage in case of errors

Bit market, easy change

Post-hoc Analysis

+ Permanent control as Black-Box-analysis

Control of the goal of the ADM System, the input, ...

Only explainable ADM systems

No ADM-system

Monopoly

Potentially high damage in case of errors.
References (sorry, German only!)

1. Brochure of the Bayerische Landesmedienanstalt Google for „BLM Dein Algorithmus - meine Meinung“

Prof. Dr. Katharina A. Zweig
zweig@cs.uni-kl.de
@nettwerkerin bei Twitter

2. Study for the Bertelsmann foundation (2018)